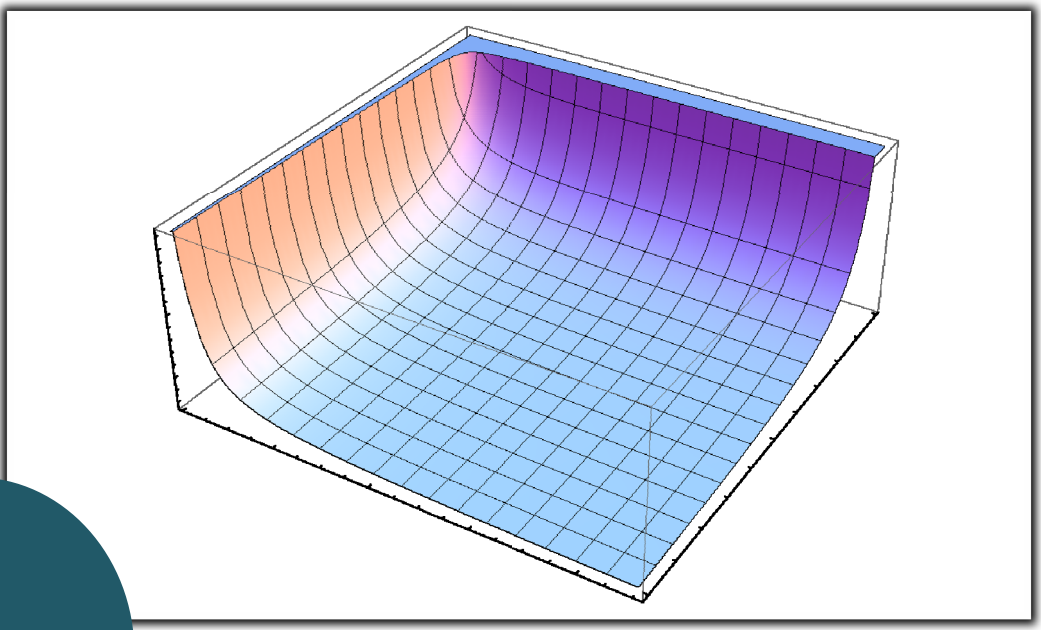
PhD Dissertation

Said Zaghloul

Design and Performance Optimization of Authentication, Authorization, and Accounting (AAA) Systems in Mobile Telecommunications Networks



Technical University of Braunschweig

Design and Performance Optimization of Authentication, Authorization, and Accounting (AAA) Systems in Mobile Telecommunications Networks

Design und Leistungs-Optimierung des AAA-Systems (Authentifizierung, Autorisierung und Abrechnung) in Mobiltelekommunikationsnetzen

Von der Fakultät für Elekrotechnik, Informationstechnik, Physik
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung der Würde eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigte Dissertation

eingereicht am:	07.Dec.2009
mündliche Prüfung am:	27.Apr.2010
von:	Said Ismail Said Zaghloul
aus:	Al Khobar
Referent:	Prof. Dipl.-Ing. Dr. techn Admela Jukan
Referent:	Prof. Dr.-Ing. Adam Wolisz
Vorsitzender:	Prof. Dr.-Ing. Rolf Ernst

Braunschweig, im May 2010

In memory of my aunt, Mrs. Yosra Katato,

Abstract

Authentication, Authorization, and Accounting (AAA) systems have been and will continue to be pivotal elements for the success of mobile telecommunications networks. In their basic operation, AAA systems grant users the required access and facilitate the collection of accounting records which reflect the subscribers' usage of network resources. The design of AAA systems is therefore instrumental to the operators' revenue growth as it largely depends on ensuring transparent verification of users' identities, quickly authorizing the requested QoS levels by the services, and implementing smart charging and accounting strategies for the supported services. In light of these developments, in this thesis, we lay the foundations for the first formal framework for AAA system planning by extending fundamental results from cellular performance studies using basic principles of probability, renewal theory, and transient Markov chains. The developed planning models estimate the AAA signaling load in mobile networks for centralized and distributed AAA topologies under a wide spectrum of design variables including protocol settings, network configuration, session statistics, and mobility profiles.

Armed with the understanding of the AAA signaling load, we designed novel optimization mechanisms for accounting and authentication. First, we developed a mechanism which keeps tight control on the reliability of the accounting process for the supported services in the mobile network. This is achieved by accounting policies which optimally resolve the trade off between the frequency of the usage reports sent to the AAA system by metering access gateways in the network and the potential revenue losses due to unreported usage if access gateways fail. Second, we proposed a novel proactive signaling mechanism which mitigates the authorization delay by utilizing the AAA framework as a bridging element between the service tier and and the radio network. Our mechanism is motivated by the fact that multiple services may be offered by third parties and across different operators and hence the resource authorization delay by services may vary during handoff events.

As AAA systems continue to proliferate beyond traditional mobile networks, we introduced AAA protocols to two promising areas including cellular backhaul applications over wireless mesh networks and inter-operator layer 2 optical communications. The thesis results illustrate the applicability of the planning models to a wide range of design

scenarios, and the scalability and robustness of the proposed AAA optimization mechanisms and applications. As such the obtained results demonstrate potential for the proposed mechanisms and solutions towards standardization and commercial products.

Kurzfassung

Systeme zur Authentifizierung, Autorisierung und Abrechnung (AAA) waren und werden auch in Zukunft entscheidende Faktoren für den Erfolg mobiler Telekommunikationsnetze sein. In ihrer grundlegenden Funktionsweise erlauben AAA-Systeme den Teilnehmern den notwendigen Zugriff auf die Netz-Infrastruktur und ermöglichen die Generierung von Abrechnungs- beziehungsweise Gebühren-Datensätzen, welche die Nutzung der Netzwerk-Ressourcen durch den Benutzer widerspiegeln. Diese Funktionalität begründet ihre Relevanz in der Migration von einfachen Daten-Verbindungsmodellen auf differenzierte Multimediadienste mit hohem Funktionsumfang sowie mobilen Breitband-Anwendungen. Das Design von AAA-Systemen ist somit unverzichtbar für die Gewinn-Maximierung des Betreibers, die stark abhängig ist von der Bereitstellung transparenter Verfahren zur Überpruefung der Nutzeridentität, schneller Autorisierung der durch die Dienste angeforderten QoS-Parameter, und der Bereitstellung intelligenter Abrechnungs-Strategien für diese Dienste. Vor dem Hintergrund dieser Entwicklung schafft die vorliegende Arbeit die Grundlagen eines ersten formalen Rahmenwerkes zur Planung von AAA-Systemen. Dies erfolgt durch eine Erweiterung fundamentaler Ergebnisse aus dem Bereich der Performance-Analyse zellularer Mobilfunknetze, sowie unter Verwendung grundlegender Prinzipien der Wahrscheinlichkeits- und Erneuerungstheorie und transienter Markov-Ketten. Das entwickelte Planungsmodell ermittelt die aufgrund der AAA-Signalisierung im Kommunikationsnetz entstehende Last, sowohl für zentralisierte als auch für verteilte AAA-Topologien unter Berücksichtigung eines breiten Spektrums von Designvariablen wie Protokolleinstellungen, Netzwerkkonfiguration, statistische Beschreibungen der Verbindungsdauern und Mobilitätsprofilen.

Basierend auf der Kenntnis der AAA-Signalisierungslast wurden Optimierungsmechanismen für die Abrechnung und Authentifizierung entwickelt. So wurde ein neues Verfahren entworfen, welches eine präzise Kontrolle der Zuverlässigkeit des Abrechnungsprozesses der in Mobilfunknetzen unterstützten Dienste ermöglicht. Dies wurde durch ein regelbasiertes Abrechnungsverfahren erreicht, das eine optimale Balance zwischen der Häufigkeit herstellt, mit der Nutzungsdaten von Zugangs-Gateways an das AAA-System gesendet werden, und einem möglichen Einnahmeverlust, der durch eine nicht

erfasste Netzbenutzung aufgrund eines Gateway- Ausfall entsteht. Weiterhin wurde ein neuer proaktiver Signalisierungsmechanismus zur Reduzierung der Autorisierungsverzögerung vorgeschlagen, der das AAA-Framework als Brücke zwischen der Service-Schicht und dem Mobilfunknetz verwendet. Dieses Verfahren ist insbesondere motiviert durch die Tatsache, dass viele Dienste von Drittanbietern gegebenenfalls sogar über Providergrenzen hinweg bereitgestellt werden, was dazu führt, dass die Autorisierungsverzögerungen der dienstspezifischen Ressourcen im Falle eines Handovers nicht vorhersagbar variieren.

Da die Bedeutung von AAA-Systeme kontinuierlich über den Bereich traditioneller Mobilfunknetze hinaus anwächst, wird die Anwendung von AAA-Protokollen in zwei viel versprechende Bereiche vorgestellt: die Anbindung von Basisstationen über vermaschte drahtlose Netze, auch als Backhaul-Anwendung bezeichnet, sowie Inter-Provider Kopplungen mittels optischer Verbindungen auf Layer-2. Die Ergebnisse der vorliegenden Arbeit zeigen, dass dieses Modell für einen großen Bereich von Anwendungsszenarien geeignet ist und demonstriert die Skalierbarkeit und Robustheit des vorgestellten AAA-Optimierungsmechanismus. Die gewonnenen Ergebnisse offenbaren sowohl ein erhebliches Potential für eine zukünftige Forschung auf diesem Gebiet, als auch potentielle Mechanismen und Lösungen hinsichtlich einer möglichen Standardisierung und Markteinführung.

Acknowledgments

The conception and fruition of this thesis work would not have been possible without the inspiration and encouragement of many people. I would like to start by expressing my sincere gratitude to my PhD advisor, Prof. Admela Jukan, for her remarkable supervision and invaluable feedback. Her sincere devotion to her research and supporting attitude to her students were genuine traits that turned my PhD experience with her fruitful and worthwhile. She sets the bar for professionals who constantly seek originality and inspire success among others. I was very fortunate to have worked with her especially for the opportunities to lead the wireless research activities in her group, present my findings in international scientific conferences, and learn from her visionary thinking as a woman scientist. I would also like to thank Prof. Adam Wolisz who kindly accepted to be my second examiner and for his valuable time to evaluate my work. I am also grateful to Prof. Rolf Ernst who generously accepted to chair the examination committee and who was always welcoming and supportive when I joined the Institute of Computer and Network Engineering (IDA).

I would also like to thank Dr. Wolfgang Bziuk, who was a very supportive colleague and a great research partner. I always enjoyed writing scientific articles with him and was fortunate to learn from his solid analytical experience. I would also like to thank Dr. Wesam Alanqar for being a wonderful research collaborator and a dependable source for professional advice. A very big thank you goes to my colleagues Mohit Chamania, Orawan Tipmongkolsilp, Xiaomin Chen, Dr. Silvana Greco, and Prof. Wael Adi for being wonderful research partners and great friends. Very special thanks to my friends Dennis Gebbers and Marcel Caria for helping me in the German translation of the thesis title and abstract. Many thanks go to our administrative staff, especially to Ms. Nadine Becker and Ms. Bettina Boettger, who made every effort to facilitate the submission of my thesis and my defense. Special thanks also go to Prof. Ernst's group for being wonderful friends and supportive colleagues.

I would also like to thank my former colleagues from Sprint who supported me while at Sprint and during my research. Special thanks go to Mr. Brent Scott, Dr. Lei Zhu, Mr. Jeremy Breau, Ms. Leyla Celebi, Ms. Preethi Prabhakar, and Ms. Sudha Sivashanmugam. Also a very big thank you goes to my friend Ehab Sultan from Broadcom for

helping me with technical details relevant to EVDO and UMTS networks. I would also like to thank Dr. Stefano Baroni and Mr. Brant Carson from McKinsey & Company for their valuable time and career support.

Special thanks also go to my former professor for their remarkable teaching and valuable time in thesis and project advising. Many thanks go to Prof. Victor Frost, Prof. K. Sam Shanmugan, and Prof. Gary Minden from the University of Kansas for their great courses and projects in networking and wireless communications. I would also like to thank Prof. Jamal Rahhal and Prof. Ibrahim Mansour from the University of Jordan for their great courses in communications and signal processing and remarkable advising.

Words will always be inadequate to thank my family for their constant support and encouragement during my research journey. Their motivating comments were always great sources for enthusiasm and achievement. Last but not least, I would also like to thank my friends Fadi Abu Shahla, Adam Bittlingmayer, Jennifer Hamdi, Abdulmueen Al Bawwab, Dr. Osamah Badarneh, and Zeid Munir for their time and for doing every possible thing to encourage me during my studies. I will always be indebted to them. For those whom I did not mention and made my work possible, thank you very much.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Authentication, Authorization, and Accounting (AAA) systems have been and will continue to be pivotal to the success of current and emerging mobile telecommunications networks. In their basic operation, AAA systems play a crucial role in granting users the required access and in facilitating the collection of accounting information which reflect the subscribers' usage of the network's resources. The design of AAA systems is therefore crucial to the operators' revenue growth as it largely depends on ensuring transparent verification of users' identities, quickly authorizing the requested QoS levels by the services, and implementing smart charging and accounting strategies for the supported services. Even further on the horizon, the reliance on AAA systems will increase as operators migrate from basic data connectivity models towards differentiated multimedia rich services and broadband mobile applications. Such vision directly translates into demands of seamless access, mobility, and guaranteed QoS for services which are all undoubtedly impacting to the design of AAA systems in the operators' network. In light of these developments and given the current design practices based on large overprovisioning, it is difficult to match the imminent customers' requirements to AAA system size and protocol settings. While it is needless to say that over-provisioning is inefficient and can be quite costly to operators, the ramifications of under-provisioned AAA systems underly intolerable risks of blocking users from access and losing valuable accounting information. The natural question that arise in this context is why can not we utilize current network design approaches to plan AAA systems ?

Current network design approaches can not be directly applied to plan AAA systems due to considerations of AAA protocol, mobility, and service statistics. Since such models are typically based on fitting traffic traces (e.g., [1–3]), they can not easily incorporate AAA protocol procedures and settings such as the frequencies of reauthentications and accounting reports from the network. This indeed limits our comprehension of the signaling protocol behavior and restricts our design means to qualitative descriptions. Furthermore, current design models typically address fixed networks and therefore do not capture effects of handoffs in mobile systems. The signaling rate towards the AAA system is largely increased when users make frequent movements or make cell reselections between base stations that belong to different "access gateways". Access gateways are analogous to routers, and base stations are similar to access points in WiFi networks. In fact, ignoring mobility can result in significant under provisioning of the system [4]. For instance, even under conservative assumptions that only 10% of the mobile users

engage in movements and cellular reselections between access gateways, the system can be easily under provisioned by more than 60% if mobility is ignored. Due to technological and demographic factors, the access gateway sizes are likely to be different, therefore, increasing the challenge of planning AAA systems due to the dependence on handoff frequencies and mobility patterns between access gateway regions. As mobile operators go towards differentiated services and broadband content, the session duration statistics will deviate from short voice calls and limited web browsing to longer sessions with significant usage resulting in larger signaling traffic towards the AAA system in the network. Thus, it is the interplay of AAA protocol procedures, mobility, and session statistics, that makes the AAA system planning problem intriguing and unique. This interplay is especially challenging when operators shift from centralized AAA deployments to multiple AAA sites for load balancing, redundancy, and signaling delay minimization. In this thesis, we fill this fundamental gap and develop the first AAA system planning models for fixed and mobile networks.

While AAA system planning addresses issues at the design time, optimizations are needed to enhance the performance of AAA systems once they are in operation. As operators migrate into mobile ecosystems supporting multiple services, the signaling rate towards the AAA system is expected to grow significantly as each service flow generates its own accounting messages such as in WiMAX and Long Term Evolution (LTE) systems [4, 5]. This clearly necessitates optimizations to accounting signaling as accidental access gateway failures can lead to unacceptable losses in terms of revenue and subscriber usage details. While such risks can be minimized by increasing the frequency of accounting reports, the signaling load on the AAA system may grow significantly leading to higher likelihoods of AAA system overload and failovers in the network [6, 7]. While suboptimal design methods based on trial-and-error and over-provisioning may be used to select suitable reporting frequencies for accounting in simple deployments, they turn quickly inefficient and cumbersome in multi-service deployments as a reporting frequency is needed for each service. Even worse, the likely change of mobility and service session statistics over time, imposes recurrent costs for tuning and upgrading the AAA system and hence reducing the perceived average revenue per user (ARPU).

Current AAA optimization literature [8–10] primarily focuses on the minimization of authentication and authorization signaling delay and load during handoff between access gateways as they reflect on the perceived QoS by the users. However, relatively much less efforts focused on optimizing aspects related to the accounting signaling traffic. In this thesis, we take up the challenge of optimizing the accounting process by developing the first formal framework that quantifies the trade off between the potential revenue loss and the signaling load in multi-service mobile networks. In addition, we contribute to the on-going efforts of mitigating the authorization delay, by designing a simple, application-layer proactive signaling mechanism which mitigates authorization latency in multi-service network deployments by considering the authorization delay requirements for each service.

As wireless systems continue to evolve, we envision that AAA applications will extend

beyond traditional mobile telecommunication networks to serve new network types such as Adhoc [11], vehicular [12, 13], cognitive radio [14], and layer 2 optical networks [15]. For instance, researchers proposed that mobile nodes extend current AAA protocol methods and act as proxies to facilitate authentication in Adhoc networks [11] and that AAA signaling be used to facilitate realtime authorization of spectrum resources in cognitive radio networks [14]. In this thesis, we unveil the potential for two novel applications for AAA systems: one relevant to cellular backhaul (i.e., the connection between base stations and radio controllers) over wireless mesh networks, and another relevant to securing inter-operator layer 2 communications in optical networks. On the one hand, our AAA application for cellular backhaul is largely motivated by the significant attention that cellular backhaul has recently received especially over wireless mesh deployments [16]. On the other hand, our AAA application for optical networks is the first endeavor to secure path computation and provisioning. It accounts for inter-operator communications in carrier grade transport technologies for optical networks.

Finally, while we endeavored in this thesis to formally address some aspects of AAA system planning and optimization issues as well as proposed new applications in two promising directions, the future is yet to uncover more applications for AAA systems accompanied with new design challenges. This process continues as the mobile communications and networking disciplines keep evolving and as broadband applications become integral parts of our daily lives.

1.1 Thesis Contributions

Our work in this thesis belongs to the general research area of control plane signaling in mobile networks. The contributions of this thesis fall into three primary fields: system planning, performance optimization, and protocol design. Relevant to the *first*, our work is the first attempt to lay generic foundations for AAA system planning in fixed and mobile networks. Our work addresses the challenging interplay among relevant design parameters including AAA protocol procedures, mobility, and session statistics. For centralized AAA system deployments, we utilize basic principles of probability and renewal theories to combine cellular performance concepts of residence and channel holding durations [17, 18] with AAA protocol operational aspects. For distributed AAA deployments, since the signaling rate may depend on the mobility pattern between access gateway regions in the network, we extend our analysis for the AAA signaling rate to consider more than one AAA system in the network with arbitrary distributions of the users' base and mobility patterns between access gateway regions.

Second, relevant to performance optimization, this thesis offers the first mechanism to ever address the challenge of protecting the operators' revenue as they migrate to IP based architectures supporting multiple services. This effort attempts to fill an important research gap as the majority of the efforts in the literature (e.g., [8–10]) focus on authentication and authorization delay issues during handoffs and ignore accounting

traffic which in many cases dominate AAA signaling (6:1 ratio according to [4]). However, we also contribute to these ongoing efforts by proposing a novel mechanism to accommodate service authorization delay variations due to third party service providers and roaming [19, 20].

Third, we introduce AAA signaling to two emerging communications networks: cellular backhaul and multi-operator optical communications. For the first we design an architecture in which wireless mesh operators offer cellular backhaul services to cellular operators and charge them depending on their resource usage. For the latter, we design a novel signaling framework that secures path computation signaling and offers accounting capabilities for optical connections spanning multiple operators.

1.1.1 System Planning

Our AAA system planning research is motivated by the growing AAA signaling load based on observations in current mobile networks [4, 21] and its expected increase with the standardization of the IP Multimedia Subsystem (IMS) as the de-facto enabling framework for next generation services [22]. In such multi-service ecosystems, access gateways and AAA systems are required to provide extensive support highly dynamic per-subscriber and per-service AAA signaling to ensure optimal network performance [23]. In addition, with the growing Mobile Virtual Network Operators (MVNO) market where third parties can offer mobile services without owning wireless network infrastructure [24, 25], the AAA signaling pertaining to MVNO users is expected to increase appreciably. For instance, according to [26, 27], the MVNO market is expected to grow from $4 billion in 2005 to top $25 billion by 2012 covering hundreds of millions of users; also the telemetry machine-to-machine market that can be facilitated through MVNO models is expected to grow from $15 billion in 2008 to $57 billion in 2014. To accommodate these recent progressions, where there is no doubt that AAA system planning is an absolute necessity, this thesis offers the first comprehensive discussion on AAA system planning ranging from centralized AAA systems in fixed networks to distributed AAA deployments in emerging mobile architectures.

Our AAA system planning discussion is novel from scope and solution perspectives. From a scope perspective, although numerous studies have addressed AAA systems from different angles, no existing work has fundamentally covered the problem of AAA system planning so far. For instance, [28] provided analysis relevant to the performance of Mobile IP Regional Registration and provided results relevant to the handover latency when contacting a central AAA system. In [29, 30], the authors provided basic analysis on the optimization signaling cost and fault tolerance for AAA systems in hierarchical Mobile IP deployments. However, common to [28–30], accounting signaling and session statistics were not considered and mobility was only addressed in simple network structures and under uniform user concentration assumptions. Many other studies which also exist in the literature addressed several other AAA system aspects including

authentication delay optimization [31, 32], the processing load estimation for ciphering algorithms [33], third party AAA applications [34], and management extensions for AAA protocols [35]. Clearly, the current literature leaves the subject of AAA system planning open for research and therefore one of our major goals in this thesis is to address this challenging aspect.

From a solution methodology perspective, the proposed planning models build on results from cellular performance and location management studies. In this regard, we utilize concepts of the residence time and call duration from call performance research [36, 37] to estimate the impact of mobility on the AAA signaling load towards a centralized AAA system based on AAA protocol procedures and settings. We then extend our analysis by using basic results from transient Markov chains theory and mobility concepts from location management studies, to address generic AAA system deployments. Our generic analysis allows the consideration of scenarios in which more than one AAA system is deployed in the network, different authentication protocols are used in different regions in the network, users may be non-uniformly concentrated, and access gateway sizes and mobility patterns can be arbitrarily defined. A cross product of this research is offering novel contributions to handoff modeling under generalized assumptions which can be applied beyond AAA systems to mobility management protocols such as Mobile IP and Proxy Mobile IP protocols.

It is noteworthy to state that the direct use of either call performance or location management models is not sufficient. This is because call performance models mainly focus on voice call metrics such as blocking and effective call duration and do not consider protocol specifics. They also make few simplifying assumptions of infinite network sizes and uniform mobility assumptions which prevent the consideration of roaming and distributed AAA deployments. Similarly, location management methods are interested in area crossings and mobility patterns but assume uniform area sizes and do not consider session duration statistics. The latter makes it difficult to apply methods from location management research to AAA scenarios as they ignore the interplay between the session and the residence times and focus on the mobility pattern. To sum up, this work elegantly combines and extends cellular and location management approaches to accurately address the specifics of AAA signaling in various mobility and network configuration deployments.

1.1.2 AAA System Optimization

While many operational issues can be avoided by proper system design and planning, dynamic mechanisms are needed to enhance the performance of operational systems. To address some of the operational aspects, in this thesis, we devise a dynamic mechanism which enhances the reliability of accounting records and propose a signaling scheme to mitigate the service authorization delay in multi-service environments. Our accounting optimization mechanism optimally balances the accounting signaling load as a function

of the potential loss from all services. For instance, for a typical size equipment [38], the failure of a network access server (NAS) serving 24,000 active users from 800 base stations with average session duration of 10 mins and a charge of 10 cents a minute, results in a loss of 12,000 USD when the reporting interval equals 10 mins. A reduction of the potential loss by half via reducing the reporting intervals, would result in requirements to handle about 30% more signaling load; a further loss reduction to 1,000 USD would require the signaling server capacity to go up to 314%.

Clearly, there is a tradeoff between the potential loss and the signaling load; the shorter the reporting interval the smaller the potential loss, but also the larger the signaling load and hence the required size of the AAA system. The proposed solution exploits standard AAA messages to adaptively update the accounting reporting intervals for all services according to their session duration and mobility statistics. This is especially important for the emerging fourth generation (4G) mobile networks where third-party services and broadband content are integral components of the 4G ecosystem and are created to be economically viable with shorter launching times. The proposed solution does not pose inter-operability issues as it does not require any modifications to the AAA protocols nor to the access gateways' implementations, and its implementation scope is limited within the AAA systems. In fact, our solution extends the concepts which we developed for AAA system planning to dynamically adapt to the system load and the potential monetary losses based on a novel process of estimating of session and mobility statistics. Although the problem is complex, our approach is yet generic and simple as it is not affected by whether the network is fixed or mobile or whether the AAA deployment is centralized or distributed. The results demonstrate that our mechanism is robust under various operational conditions, easy to implement, and offers considerable potential for loss control compared to the current static approaches.

In addition to accounting optimization, we extend the current body of research on authorization delay optimization (e.g., based on fast Mobile IP, hierarchical AAA designs, and Media Independent Pre-Authorization framework [8, 9, 39]) to incorporate the authorization delay which occurs at the service tier in multi-service mobile networks. Specifically, we propose a novel proactive mechanism that exploits the policy framework in emerging multi-service architectures [5, 22] to mitigate the variability of authorization delay when third party application servers are used. Our mechanism utilizes the fact that in many technologies (e.g., EVDO), the AAA system has an existing network interface to the radio network controller. Since the policy system usually uses AAA protocols, the AAA system can act as a bridging component between the radio network and the policy system. Thus, service authorization delay estimates are passed from the policy system in the service tier through the AAA system to the radio network controller and proactive triggers are sent out to the policy system through the AAA system. As such, our mechanism enhances the session continuity likelihoods as users move between different access gateways in the network and as they roam from one operator into another. The results are promising as they demonstrate the ability of our proactive scheme to mitigate service authorization delay and show that our scheme scales similarly to the original system as function of many design variables.

1.1.3 New Applications

Inspired by the success and the flexibility of AAA systems to support emerging mobile networks as well as by their use in non-traditional networks such as adhoc, mesh, and cognitive radio systems, we explore their applications to two novel applications: in cellular backhaul applications over wireless mesh networks and in multi-operator layer 2 optical networks. Relevant to the first, we propose the first billing architecture [20] for cellular backhaul applications over wireless mesh networks and analyze its scalability. While this is only the first step to address the generic area of billing for multi-service wireless mesh networks, when applied to cellular backhaul, it poses a few practical and new challenges that have not been studied so far. First, adding or releasing backhaul bandwidth chunks reflects heavily on signaling for billing updates. Second, the performance of every billing scheme highly relates to the dynamic bandwidth reservation mechanisms at the base station. It is very critical that the billing signaling traffic resulting from bandwidth reservations is minimal in order to ensure scalability for the billing architecture. Using threshold based bandwidth reservation policy, the results are promising and show that our billing scheme scales well even for poor implementations of bandwidth reservations.

Second, we also propose a novel AAA signaling framework for inter-carrier path provisioning in carrier grade optical networks. Our application is largely motivated by the evolution of carrier grade transport technologies, most notably based on the Ethernet and MPLS paradigms, which were catalyzed by the ever-growing demands of broadband applications for science and entertainment. Our AAA signaling method addresses the lack of security and accounting features within the path computation architecture and concentrates on aspects of multi-domain routing and path provisioning. The proposed mechanism is shown to facilitate secure exchange of path computation signaling among domains, associate path setup with the computed paths, while enabling sharing of accounting information between carriers. The illustrative results show that the proposed signaling framework is lightweight and scales linearly with the number of domains. The results and the reliance on universally accepted protocols in our AAA framework clearly demonstrate its promising potential for seamless integration within the path computation platforms and largely facilitate its implementation in commercial deployments.

1.2 Supporting Publications

1.2.1 Journal Articles and Book Chapters

1. S. Zaghloul, A. Jukan, “Optimal Accounting Policies for Reliability and Capacity of AAA Systems in Mobile Networks,” to appear in the IEEE Transactions on Mobile Computing Journal, Jun 2010

2. O. Tipmongkolsilp, S. Zaghloul and A. Jukan, "The Evolution of Cellular Backhaul Technologies: Current Issues and Future Trends," to appear in the IEEE Communications Surveys & Tutorials Journal, 1st Quarter 2011
3. W. Bziuk, S. Zaghloul and A. Jukan, "A New Framework for Characterizing the Number of Handoffs in Cellular Networks," European Transactions on Telecommunications, 20(7), pp. 689-700, Sep 2009
4. S. Zaghloul, A. Jukan, "Signaling Rate and Performance for Authentication, Authorization, and Accounting (AAA) Systems in all-IP Cellular Networks," IEEE Transactions on Wireless Communications, 8(6), pp. 2960-2971, Jun 2009
5. S. Zaghloul, W. Bziuk, A. Jukan, "Signaling and Handoff Rates at the Policy Control Function (PCF) in IP Multimedia Subsystem (IMS)," IEEE Communications Letters Journal, 12(7), pp. 526-528, Jul 2008
6. S. Zaghloul, A. Jukan, "Architecture and Protocols for Authentication, Authorization and Accounting (AAA) in the Future Wireless Communications Networks," Handbook of Research on Wireless Security, Chapter XII, Auerbach Publications, Taylor & Francis Group, pp. 158-175, 2008, ISBN: 159904899X
7. S. Zaghloul, A. Jukan, W. Alanqar, "Extending QoS from Radio Access to all-IP Core in 3G Networks - An Operator's Perspective," IEEE Communications Magazine, 45(9), pp. 124-132, Sep 2007
8. S. Zaghloul, A. Jukan, "Relating the AAA and the Radio Access Rates in 3G Cellular Networks," IEEE Communications Letters Journal, 11(4), pp. 363-365, Apr 2007
9. M. Chamania, S. Zaghloul, S. Greco Polito and A. Jukan, "Towards a Scalable AAA Signaling for Inter-carrier PCE Framework," revision submitted to IEEE Communications Magazine, 2009
10. S. Zaghloul, A. Jukan, "A Generic Framework for Planning AAA Signaling in Centralized and Distributed Network Deployments," under journal submission
11. W. Bziuk, S. Zaghloul and A. Jukan, "Revisiting the Handoff Statistics Theory for Next-Generation Wireless Cellular Networks," under journal submission

1.2.2 Conferences and Workshops

1. S. Zaghloul, W. Bziuk and A. Jukan, "A Novel Analytical Framework for Mobility Modeling in All-IP Wireless Systems," proceedings of the 21st International Teletraffic Congress (ITC 21) proceedings, Paris, Sep 2009
2. P. Pereira, S. Zaghloul, A. Jukan, S. Glaeser, "Towards Innovative Vehicle to Application Server Communications: An IMS Centric Approach," proceedings of the 4th ACM International Workshop on Mobility in the Evolving Internet Architecture (MobiArch'09), Krakow, Poland, Jun, 2009

3. W. Bziuk, S. Zaghloul and A. Jukan, “The Spatial Effect of Mobility on the Mean Number of Handoffs: A New Theoretical Result,” proceedings of the IEEE International Communications Conference (ICC’09), Dresden, Germany, Jun 2009
4. S. Zaghloul, J. Aznar and A. Jukan, “Application Layer Signaling for Proactive Handoff Management in all-IP Wireless Networks,” proceedings of the IEEE International Communications Conference (ICC’09), Dresden, Germany, Jun 2009
5. W. Bziuk, S. Zaghloul, A. Jukan, “A New Framework for Characterizing the Number of Handoffs in Cellular Networks," the Fifth Polish-German Teletraffic Symposium (PGTS 2008), Berlin, Germany, Oct 2008
6. S. Zaghloul, W. Bziuk and A. Jukan, “A Scalable Billing Architecture for Future Wireless Mesh Backhauls,” proceedings of the IEEE International Communications Conference (ICC’08), Beijing, China, pp. 2974-2978, Jun 2008
7. S. Zaghloul, A. Jukan, “A Simple Signaling Mechanism for Seamless Inter-operator Mobility in All-IP Networks,” proceedings of the 5th Annual IEEE Consumer Communications & Networking Conference (CCNC’08), Las Vegas, USA, pp. 381-385, Jan 2008
8. S. Zaghloul, A. Jukan, “On the Performance of the AAA Systems in 3G Cellular Networks,” proceedings of the IEEE International Communications Conference (ICC’07), Glasgow, United Kingdom, pp. 2103-2108, Jun 2007

1.3 Thesis Organization

This thesis is structured in six Chapters. Following the Introduction, Chapter 2 provides an overview of AAA systems and their applications in mobile networks. Chapter 3, describes the AAA system planning framework. It starts by modeling the AAA signaling load in estimation fixed networks. It then generalizes the analysis to mobile environments for centralized and distributed AAA system deployments including roaming. We conclude Chapter 3 with advanced discussion of our novel contributions to handoff modeling which is relevant to the AAA planning process. Chapter 4 develops AAA optimization methods and describes novel AAA applications. It starts by describing the optimization approach for mitigating the authentication delay in multi-service architectures followed by the accounting optimization framework. It then illustrates further applications of AAA protocols in the areas of cellular backhaul over wireless mesh networks and layer 2 inter-operator optical communications. In Chapter 5, we show results that quantify the impact of the proposed concepts relevant to AAA system planning and performance optimizations. We dedicate Chapter 6 for Conclusions and directions for further work.

Chapter 2 Overview of AAA Systems

Authentication, Authorization, and Accounting (AAA) systems play an instrumental role to the success of service delivery. Typically, users are authenticated when requesting a service and only after successful authentication they are authorized to use the service. Once the user is granted access to the service, the network generates accounting messages based on the user's activity.

Currently, the Remote Authentication Dial In User Service (RADIUS) protocol [40, 41] is the most widely deployed AAA protocol in cellular networks such as LTE, EVolution Data Optimized (EVDO), and WiMAX [42–44]. Due to its inherent security and reliability weaknesses, RADIUS is to be substituted by Diameter [45], in the upcoming years. Therefore, fourth generation cellular standards such as Long Term Evolution (LTE) rely on the Diameter protocol for AAA signaling especially for the invocation and control of value added services. In this chapter, we provide a short overview of AAA standards and their applications within cellular networks. We conclude our discussion with a brief summary of the current research in the area of AAA architectures

2.1 Supporting Publications

1. P. Pereira, S. Zaghloul, A. Jukan, S. Glaeser, "Towards Innovative Vehicle to Application Server Communications: An IMS Centric Approach," The 4th ACM International Workshop on Mobility in the Evolving Internet Architecture (MobiArch'09), Kraków, Poland, Jun 2009
2. S. Zaghloul, A. Jukan, "Architecture and Protocols for Authentication, Authorization and Accounting (AAA) in the Future Wireless Communications Networks," Handbook of Research on Wireless Security, Auerbach Publications, Taylor & Francis Group, Chapter XII, pp. 158-175, Mar 2008, ISBN: 159904899X
3. S. Zaghloul, A. Jukan, W. Alanqar, "Extending QoS from Radio Access to all-IP Core in 3G Networks - An Operator's Perspective," IEEE Communications Magazine, 45(9), pp. 124-132, Sep 2007,

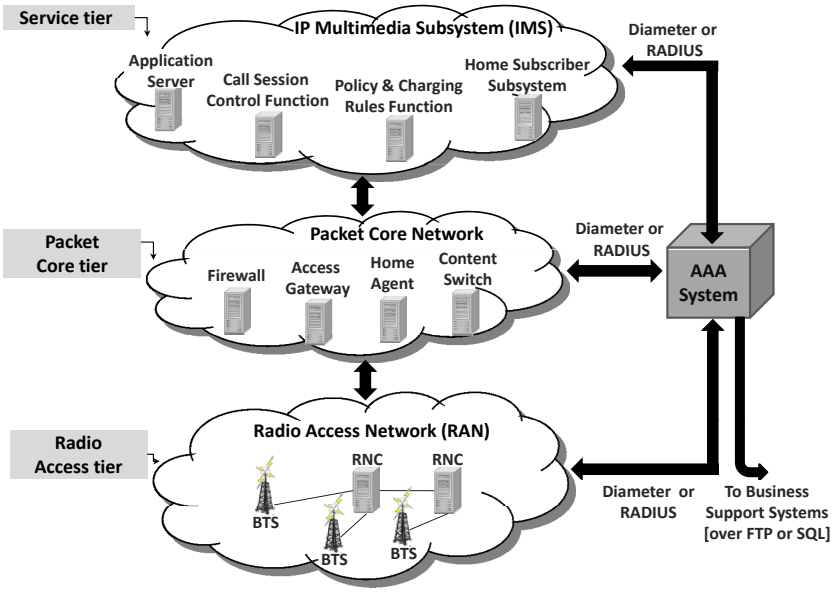


Figure 2.1: A simplified cellular architecture with access, core, and service tiers. [RNC: Radio Network Controller, FTP: File Transfer Protocol, SQL: Structured Query Language]

2.2 Introduction to AAA Systems in Mobile Environments

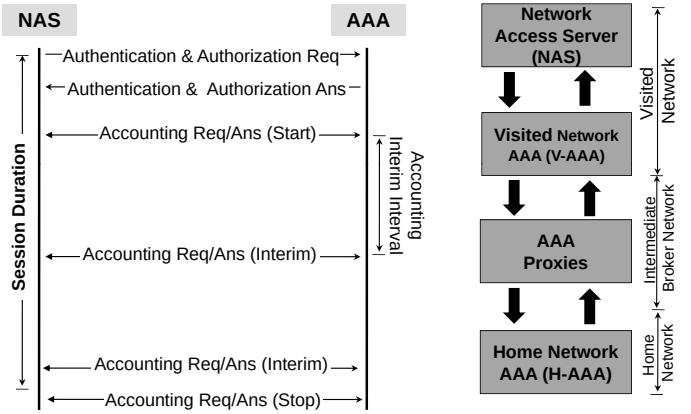
We start our introduction by visualizing the mobile network as three logical tiers: radio access, packet core, and service tiers. As shown in Fig. 2.1, the Radio Access Network (RAN) consists of the Base Transceiver Stations (BTSs) and the Radio Network Controllers (RNCs) which compose the access network and offer wireless connectivity to the users, and management and control functionalities such as radio resource, mobility, and admission control. The access network has an interface with the AAA server in some systems (e.g, WiFi 802.11 and EVDO) in order to authorize the users' access and establish air link security. The core tier includes elements that support IP connectivity and transport, firewalls, Dynamic Host Configuration Protocol (DHCP) servers for dynamic IP addresses, home agents to maintain users' IP addresses as they move between IP gateways in the network, and content switches. The core tier also includes an important element called the Access Gateway (AGW) which generically refers to the first IP gateway for users' traffic. Examples of AGWs include the Packet Data Serving Node (PDSN) in EVDO systems, the Access Serving Node Gateway (ASN-GW) in WiMAX systems, the Gateway GPRS Support Node (GGSN) in UMTS, and the

Serving Gateways (S-GW) and Packet Data Network Gateways (PDN-GWs) in LTE. Such core elements connect to the AAA system to authenticate and authorize the users' access request for IP services. Once the user is authenticated, the AGW typically generates accounting messages to reflect the usage of IP services. Accounting records are forwarded by the AAA to the Business Support System (BSS) for further processing and to generate the users' bills as we will discuss in the next section.

The service tier is expected to follow the IP Multimedia Subsystem (IMS) framework [22] or similar paradigms. In IMS, services are hosted by application servers and signaled by the Session Initiation Protocol (SIP) for session establishment, modification, and tear down over a group of SIP servers known as Call Session Control Functions (CSCFs). IMS relies heavily on Diameter for authentication by allowing CSCFs and application servers to contact the users' profile database (a.k.a, the Home Subscriber Subsystem (HSS)). Diameter is also used for controlling the QoS levels for the invoked services and their corresponding charging rules using the so-called Policy and Charging Rules Function (PCRF). Finally, Diameter signaling may also be used for accounting where some CSCFs generate accounting records towards the AAA system during the lifetime of the session.

From an AAA protocol perspective, the network element that interacts with the AAA server to authenticate the users' access and/or to generate their accounting is called the Network Access Server (NAS). For example, the NAS can be a router, a RNC, an AGW, or a CSCF. When the user attempts to access the network or use a service (e.g., a VoIP session), the NAS may authenticate the user through an authentication protocol such as the Password Authentication Protocol (PAP), the Challenge Handshake Authentication Protocol (CHAP), Extensible Authentication Protocol (EAP), or HTTP based authentication. Upon obtaining the responses from the client, the NAS generates an authentication request towards the AAA server to validate the user's response. The AAA server validates the response by communicating with an external database that contains the user's credentials and authorized services and returns an access accept message if responses are valid. The access accept message may contain authorization information such as filters to grant the user access to internal networks, specific routing instructions, QoS settings, etc. This authorization set is returned as a group of Attribute Value Pairs (AVP) in the access accept message. Afterwards, the NAS may generate accounting messages based on user's activity (connection time, total bytes used, etc).

Accounting is composed of three primary accounting request and answer messages including, Start, Interim, and Stop. The accounting start message is sent at the beginning of the session and reports zero usage (e.g., zero time or capacity units). Then, accounting interim messages are sent periodically after each *accounting interim interval* passes. Accounting interim messages are used to report the accumulated usage so far and is configured by the administrator to minimize the capital loss due to the unreported usage if the NAS fails [6]. When the session terminates, an accounting stop message is sent to report the session's total resource usage. Fig. 2.2(a) shows the AAA signaling process.



(a) AAA protocol signaling

(b) AAA proxy chain in roaming scenarios

Figure 2.2: AAA systems: protocol and proxy chain operation.

Usage may be metered by the NAS on a per session (i.e., data connection) basis regardless of the invoked services. In this case accounting records only reflect the total volume and/or time used by *all* services without differentiation or indication of their specific usage. On the other hand, flow based accounting is used when it is desired that flows be differentiated. In this case, accounting records are generated for each flow separately as if each flow is running its own connection. Accounting records for all flows are correlated using one global session identifier (e.g., Session-ID) while each flow can have its sub-session identifier (e.g., Accounting-Sub-Session-Id) [45].

Sometimes an AAA server serves as client/proxy when it is provisioned with a policy instructing it to forward the request to another AAA server. Such policies are occasionally based on the domain in the user's name (a.k.a, Network Access Identifier (NAI)). Standards [46] refer to this setup as the proxy-chain configuration. For instance, in a roaming scenario the host AAA is usually configured to forward AAA requests from the hosting NAS in the visited network to the home network's AAA. Multiple proxies may be traversed along the path to the home AAA server (e.g., through broker networks which handle billing reconciliation) as shown in Fig. 2.2(b).

Although both RADIUS and Diameter generally support similar procedures, they entail critical operational differences. RADIUS runs over the intrinsically unreliable User Datagram Protocol (UDP) and generally uses a shared secret between the NAS and the AAA system to secure some important fields such as the user responses. On the other

hand, Diameter operates over reliable protocols (i.e., the Transmission Control Protocol (TCP) or the Stream Control Transmission Protocol (SCTP)). It secures its connections using proven protocols such as Transport Layer Security (TLS) or Internet Protocol Security (IPsec). Unlike RADIUS, Diameter supports stateful and stateless operation for the sessions it serves and allows server initiated requests towards its clients. It also supports a standardized fail over handling mechanism using Diameter's watchdog mechanism. Finally, Diameter is extensible and can accommodate new authentication or accounting schemes by defining Diameter applications which are negotiated during connection establishment.

2.3 Overview of Accounting

There are two major billing architectures; one for postpaid services and another for prepaid services. In postpaid systems, the users' accounts are charged offline and only the AAA system is involved in collecting accounting information and forwarding it to the Business Support System (BSS) for further processing. On the other hand, in prepaid systems, realtime charging and authorization, rating, and balance management for services is required. When the users' balance depletes, the system "hot-lines" users and redirects them to a portal where they can replenish their accounts. In prepaid systems, AAA signaling is used between the NAS which meters the service, credit control server (a.k.a, prepaid server) which manages the user's balance in realtime by interacting with the BSS, and the AAA system which authorizes access and collects accounting records.

In postpaid systems, accounting records are collected by the AAA system from all network access servers and are stored in a simple format (e.g., comma separated fields text files or SQL databases) referred to as the Usage Detail Record (UDR). The UDRs are then forwarded to the BSS using protocols like FTP or SQL [47, 48]. In the BSS, the UDRs are validated, aggregated for each session, and converted into a common format (a.k.a., normalized). The normalized session UDRs are then rated using the rating engine using the pricing plans database and are forwarded to the billing system which produces the end customer's bill. The BSS also includes other relevant systems such as Fraud Management Systems (FMS), Customer Relationship Management (CRM) systems, Order Management Systems (OMS), and any web based self service portals for the customers to manage their account or check their balance [48, 49].

In prepaid systems, credit control (prepaid) servers are introduced to allow real time rating of the requested resources and to ensure that sufficient balance exists in the users' account prior to service delivery. The credit control servers may interact with the BSS for the purposes of rating and balance management [50]. The interactions between the NAS and the AAA with the credit control server are facilitated using Diameter or RADIUS signaling [50, 51] as shown in Fig. 2.3. Credit authorization may be combined with the authentication and authorization process or delayed afterwards. The first case is used when charging for network access (e.g., establishing a data connection) and

therefore the AAA interacts with the credit control server as part of the authentication process. On the other hand, delayed credit control signaling can be used for value added services where users may invoke services after the initial registration/authentication to the network (e.g., making a VoIP call or initiating video sessions by an already authorized subscriber at different times). In this case, the credit control client (i.e., the AGW in our simplified example) interacts directly with the credit control server.

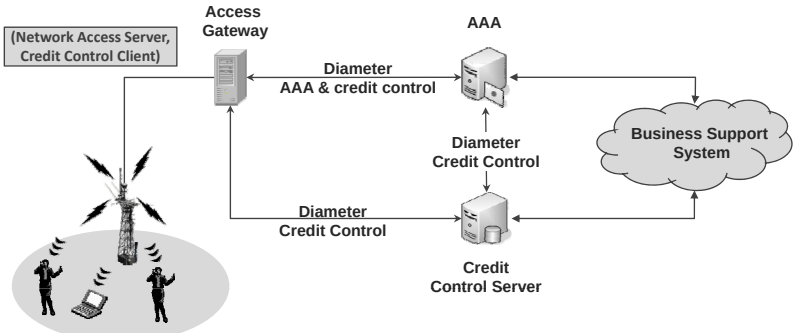


Figure 2.3: A simplified architecture for prepaid accounting systems.

In systems which offer concurrent usage of multiple services, prepaid system operation is much more complex than postpaid systems due to the realtime nature of prepaid environments. This is the main reason why services are usually first introduced to postpaid users. Credit control servers manage multiple services using credit pools. Credit pools are used to avoid issues which arise when different quotas are assigned for each service in multi-service environments. For instance, say that quota 1 and 2 share the same credit pool and that quota 1 represents 3 MB of usage and quota 2 represents 5 MB for another service. Then, the authorized pool has 8 MB worth of granted units. Now if quota 1 has been used up while quota 2 has plenty of credit units (say 4 MB free), the credit control client (i.e., in the AGW) can let quota 1 go over its granted units (i.e., more than 3 MB) as long as the authorized credit-pool units are sufficient. Without credit pools, whenever a user service depletes its granted quota units, credit control interrogations are generated towards the credit control and business support system resulting in a potential system overload. In addition, services with similar rates may be categorized into rating groups such that the credit control client can assign usage units to services without contacting the credit control server.

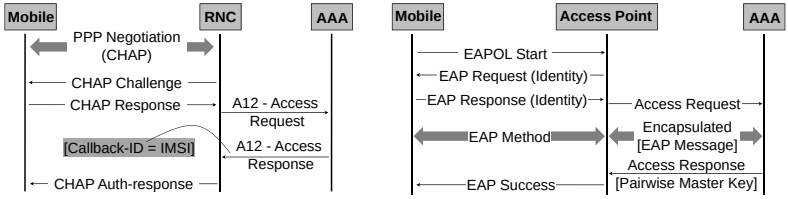
It is noteworthy to state that recent design trends in the BSS are to combine postpaid and prepaid system capabilities to support advanced service and price plans based on personal preferences [52–54]. This is referred to as *unified or converged billing*. Such unified approach combines advantages of both systems by offering postpaid customers the abilities to view their balances in realtime as well as introduce "caps" on their usage,

while at the same time reducing the cost of offering services to prepaid users [55]. For instance, in a family postpaid plan, the parents may want to assign their children a quota that once exceeded are only allowed to make calls to family members only. Similarly, prepaid users can enjoy capabilities like having some of their services covered by a prepaid account while other services such as voice are postpaid. Finally, from an AAA signaling perspective, this may require that credit control servers be used for all customers and that accounting data is forwarded immediately to the BSS for processing.

2.4 Exemplary AAA Signaling in Cellular Network Tiers

Cellular systems usually implement device authentication at the radio network tier, followed by user authentication with the AGW in the packet core tier, and finally registration authentication with the service tier.

2.4.1 In the Radio Access Tier



(a) A12 authentication signaling (b) EAP authentication signaling in WiFi

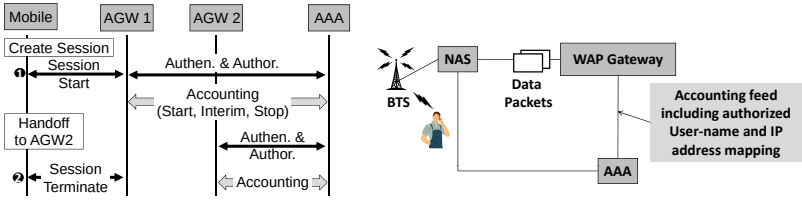
Figure 2.4: AAA signaling in the radio access tier.

Figure 2.4(a) shows a simplified EVDO network with a Radio Network Controller (RNC) and an AAA system. The interface between the RNC and the AAA is called the A12 interface as is based on RADIUS protocol [56, 57]. Mobile devices are authenticated using a CHAP based mechanism. For authenticated users, the AAA system returns the subscriber's International Mobile Subscriber Identity (IMSI) in the Callback-ID AVP to the RNC in the RADIUS access-accept message. The IMSI is used to establish a user specific connection between the RNC and the AGW (i.e., the PDSN in EVDO networks).

As systems evolve and as designers seek more universal authentication mechanisms, the 802.1x authentication mechanisms based on the Extensible Authentication Protocol (EAP) suite, used in WiFi, are being adopted by emerging systems like WiMAX and

LTE for device and user authentication. EAP is a generic framework for authentication and includes multiple authentication mechanisms, such as EAP-Transport Layer Security (EAP-TLS), EAP-Tunneled Transport Layer Security (EAP-TTLS), and EAP-Internet Key Exchange (EAP-IKE), and its generic message flow is illustrated in Fig. 2.4(b). The outcome of such exchange is usually a pairwise master session key between the access point/base station and the mobile node. This master key is then used to establish dynamic traffic and key encryption keys between the access point and the mobile node using Temporal Key Integrity Protocol (TKIP) or Counter Mode with Cipher Block Chaining Message Authentication Code Protocol (CCMP) protocols [58]. However, EAP poses a design challenge as unlike currently simpler authentication methods it results in relatively high signaling load on the airlink and on the AAA server. For instance, EAP-TTLS may result in 12 to 14 messages between the mobile node and the AAA system before the link is established.

2.4.2 In the Packet Core Tier



(a) AAA signaling by AGWs (b) AAA accounting feed for WAP gateways

Figure 2.5: AAA signaling in the packet core tier.

AAA signaling is used in the packet core tier to support mobility management between AGWs using protocols such as Mobile IP and Proxy Mobile IP versions 4 and 6 [59–64]. Mobility management is needed to maintain IP connectivity as mobile nodes move between cells served by different AGWs. However, users are authenticated as they move between AGWs and accounting signaling is generated to reflect the change of the serving AGW. In some systems (e.g., EVDO [44]), mobility between base stations or even sectors may also result in having the AGWs generate accounting records reflecting the change in the base station identifiers. This is especially useful for location based services and legal interception applications. Thus, mobility can largely increase accounting signaling and complicate correlating accounting records in the business support systems [4]. As shown in Fig. 2.5(a), when the user starts a session, AAA signaling is invoked and accounting is reported by the serving AGW to the AAA system. When the user moves to another AGW area (i.e., from AGW_1 to AGW_2), an accounting stop is sent by AGW_1 and an accounting start is sent by AGW_2 starting a new accounting ses-

sion as observed by AGW_2. Accounting records generated by AGW_1 and AGW_2 may have different session identifiers and hence a more global identifier called Acct-Multi-Session-Id is used to correlate accounting records. Finally, to facilitate the correlation of accounting records in mobile networks, further attributes are used to allow the billing system to know whether the accounting records belong to a new session, a handoff session, or a terminating session as shown in Table 2.1. In Chapter 4, we use these attributes to design an optimization mechanism for accounting in mobile networks. Finally, it is noteworthy to state that AAA signaling is used for correlating user names to IP addresses in packets observed by content switches and Wireless Application Protocol (WAP) gateways by providing accounting messages which carry such information from the AAA server to such network systems as shown in Fig. 2.5(b) [65, 66].

Table 2.1: Attributes used for accounting in mobile networks [43, 44].

Request Type	First AGW	Intermediate AGWs	Last AGW
Start	Begin of Session = true	Begin of Session = false	Begin of Session = false
Stop	Session Continue = true	Session Continue = true	Session Continue = false

2.4.3 AAA Signaling in the Service Tier

Diameter based AAA signaling [45] is used in the service tier to authenticate users and to download their profiles from the Home Subscriber Subsystem (HSS) during the registration process. Diameter is also used to signal policy and charging rules from the service tier to the packet core and radio access tiers using the Policy and Charging Rules Function (PCRF). Therefore, the PCRF might be viewed as the gluing element which allows services to communicate its QoS requirements from an application specific description to network and radio specific formats. Fig. 2.6 illustrates a simplified signaling flow for user registration with the HSS and for service invocation. Notice that messages sent using Diameter are highlighted and shaded for clarity. Registration (see steps 1-6) is needed so that the users' IP addresses are mapped to their SIP identities (e.g., `alice@operator1.com`). It is also required to allow the serving CSCF to download the users' service profiles from the HSS database. The serving CSCF is the element that inspects all user's requests and confirms that they abide by access rights specified for that user. It also acts as a SIP router and determines whether the SIP message needs to be sent to one or more application servers before granting service [22].

The registration process starts at the mobile device (step 1) by initiating a registration request towards the service tier. Once the serving CSCF receives the Register message, it issues a Diameter Multimedia Authentication Request towards the HSS (step 2) to obtain appropriate authentication vectors to challenge the user. Afterwards, it formats a SIP response (401 Unauthorized) that carries a challenge to the user (step 3). Once the

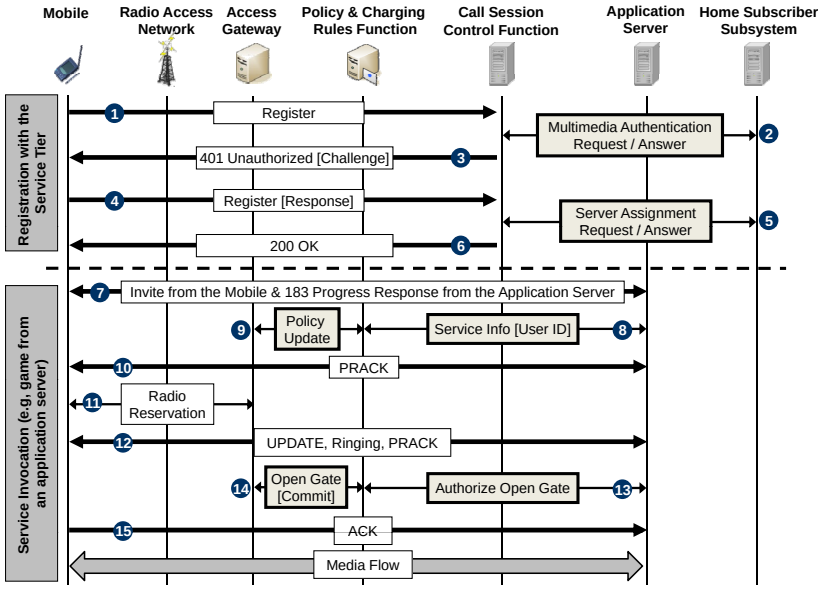


Figure 2.6: Simplified signaling for registration and service invocation in the service tier.

mobile receives the response, it immediately responds with another registration message carrying a response for the supplied challenge (step 4). When the serving CSCF receives the second registration request (Step 4), it validates the user's response. If successful, it issues a Diameter Server Assignment Request (step 5) towards the HSS requesting to be assigned for the user and querying for the user's service profile. Once it obtains the user's service profile, the serving CSCF issues a SIP 200 OK message to the mobile and hence completes the registration process. Now, the mobile is ready to initiate and receive multimedia service sessions at any time.

When the user wishes to initiate a session (step 7), it initiates a SIP Invite message towards the other end point (e.g., to a gaming application server in the network). After checking any service specific policies by the serving CSCF, the application server responds with the "183 Progress". Afterwards (step 8), the application server or the CSCF communicates the QoS information for the flow in the session description protocol (SDP) to the PCRF. The PCRF executes its pre-configured local policies, converts the authorized QoS requirements from service level description to bearer level requirements (i.e., media type, IP addresses/ports, direction, bandwidth, etc). The PCRF then (step 9) signals its rules to the AGW for the purposes of proper QoS enforcement and charging in the packet core and radio access tiers. This procedure in steps 8-9 is sometimes referred to as Service Based Bearer Control (SBBC) [67]. In step 10, a SIP

Provisional Acknowledgement (PRACK) is sent triggering QoS reservation in the radio access network (step 11). Once radio and network resources are reserved, a SIP Update message is sent to indicate reservation success (step 12). The Update message is necessary as the radio reservation may result in different media QoS settings from the ones negotiated in the SIP Invite message (step 7) possibly due to varying radio conditions. The application server acting as the other session end point responds with a SIP Ringing message and the mobile responds with a SIP PRACK (see the SIP preconditions framework in [68] for details). Depending on the operator's implementation, if the "open gate" command was not included in steps (8-9), then the application server or the CSCF instructs the AGW via the PCRF to "open the gate" for the IP flows (steps 13-14). Finally, the mobile sends a SIP ACK message for the Invite (step 15) allowing the media to flow while the AGW generates accounting records using the charging rules from the PCRF. The CSCF or the application server may also issue accounting messages pertaining to the service if desired. Once the session terminates a SIP Bye message (not shown) is sent by either the application server or the user and a corresponding accounting stop is generated accordingly.

Finally, it is noteworthy to state that Single Sign On (SSO) techniques are proposed to assure third party application servers in the form of an assertion (e.g., token) that the user is registered with his network. Thus, when a service offered by a third party is invoked, the Identity Provider (IdP) in the user's home network offers the assertion tokens to the service provider and relieves the user from re-authenticating with it. For instance, the authors in [69, 70], use an IdP such that as soon as the user is authenticated and registered to the IMS layer, the identity provider is notified using AAA signaling and hence in turn informs third parties of the authenticated user. Examples of IdPs are OpenSSO, Shibboleth, OpenID, and Microsoft CardSpace.

2.5 AAA Systems in Research

Numerous publications and standards address AAA systems from a security perspective from several angles including confidentiality, integrity, availability, authenticity, and trust and key management [71–79]. In this section, we identify promising directions in AAA research and standardization that go beyond security aspects. This includes areas of AAA system planning and scalability, signaling delay minimization, accounting optimization, and design for emerging networks (e.g., cognitive radio, adhoc, and mesh networks).

Proper AAA system planning to handle the signaling load from network access servers in the network is of significant importance. This is because under-provisioned AAA systems can result in blocking users from access and can lead to loss of revenue if accounting records are dropped. This issue is of concern as emerging cellular architectures are expected to pose a significant load on AAA systems. For example, a comparison of AAA signaling for WiMAX flow based accounting due to a session comprised of

four flows versus WiFi results in 24:1 growth of the signaling load as stated in [4]. This situation is expected to aggravate in emerging systems due to mobility between access gateways - or even between cells when accounting is enabled to report events when users move between cells/sectors - and due to the foreseen longer session durations. In addition, in large networks, AAA systems are likely to be non-centralized to load balance the control plane and to reduce authentication delay. Currently, AAA system planning models are lacking and the analytical tools for evaluating realistic design choices for mobile networks are also missing. As such, operators are only left to design their AAA systems by large over-provisioning which is not only a costly and space inefficient method, but also gives no guidelines on scalability or system settings. In Chapter 3, we fill this fundamental gap and offer AAA planning models for fixed and mobile networks in centralized and distributed AAA deployments.

The AAA signaling delay is another important metric to consider as it can result in longer session setup delay and can cause QoS degradation or session dropping during handoffs. Authentication delay during handoff received significant attention and was addressed in several efforts in the context of Mobile IP and EAP protocols [31, 39, 80–83], and most recently, in the Media Independent Pre-authentication (MPA) framework in [8]. MPA allows pre-authentication of the subscribers at the target AGW while in the source AGW region whenever a handoff is perceived to be imminent. However, the MPA framework is mostly focused on the core network tier systems[1] and does not address AAA signaling in the service tier due to QoS authorization by the Policy and Charging Rules Function (PCRF) which can be quite long (over 1 sec) depending on whether application servers need to be contacted during the authorization process. We devise a novel signaling mechanism that addresses this important issue of pre-authorization in the service tier in Chapter 4.

Optimizations for accounting schemes are important in order to control the potential loss in the event of NAS failures and to minimize the likelihood that on-going sessions are force terminated due to account depletion. Accounting reliability is most relevant to postpaid systems where accounting records are the only source of information about the subscribers' usage and their loss directly results in loss of revenue. In this case, the accounting interim intervals should be chosen such that the potential loss in the event of NAS failure is minimized while at the same time the AAA system is not overwhelmed by an excessive number of accounting requests. This problem was described in [7], however, only qualitative discussion was provided. In Chapter 4, we propose an adaptive optimization mechanism that mitigates this problem in mobile networks supporting multiple services.

For the issue of on-going sessions disruption due to balance depletion, it is important for such systems to choose a proper recharge threshold such that the user is reminded before the account is depleted. This threshold should not be too small as some active sessions may be force terminated. Optimal threshold solutions were proposed in [84–

[1] Handoff delay in the radio tier is insignificant (40-70ms) and is usually handled by radio technology specific protocols.

86] such that services are not admitted when the account balance falls below a certain threshold. Suitable times are also given such that users are notified to replenish their accounts before service termination. Furthermore, the signaling load to the credit control server due to QoS updates in the network was studied in [87] and an optimization method that trades off the accuracy of the balance and the signaling load was offered as a work around. However, we believe that work on optimal threshold selection methods for emerging multi-service environments with mobility is still preliminary and requires deeper investigation and is thus part of the future work for this thesis.

Finally, significant research efforts are underway to integrate AAA capabilities in emerging networks and architectures such as wireless adhoc networks, vehicular networks, cognitive radio systems[2], and for layer 2 optical networks [11, 12, 14, 15, 78, 79, 88, 89]. For instance, in [11] AAA solutions were proposed for adhoc networks such that mobile nodes act as auxiliary authenticators (i.e., EAP relays) to the access point which acts as the primary authenticator. In addition, researchers proposed virtual money concepts in vehicular networks where vehicles collect points for forwarding packets to other vehicles. Reputation based approaches were also proposed such that the forwarding behavior of vehicles is observed to establish trusted routing paths [89]. In addition, AAA signaling was also used to support innovative car diagnostic services where interim records were used to establish a record of operations performed with the vehicle (e.g., update flash, electric control unit reset, etc) for the sake of auditing and billing [13].

Furthermore, AAA signaling was proposed to cognitive radio systems to allow dynamic and realtime authorization of spectrum resources without the need to go through lengthy governmental regulations [14]. Moreover, AAA solutions were recently proposed in [15] to secure path computation signaling in layer 2 inter-carrier optical communications for broadband commercial and scientific applications with QoS guarantees. In Chapter 4 in this thesis, we extend the work in [15] by designing AAA flows for path setup and reservation. We design a multi-domain accounting scheme for the participating domains based on concepts from cellular networks. Finally, we also propose further applications to AAA signaling for cellular backhaul applications. Cellular backhaul refers to the links between the base station sites and the radio network controllers in the core network. In our work, we propose a scalable accounting mechanism for cellular backhaul connections when running over wireless mesh networks owned by third party providers.

[2]In cognitive systems, licensed spectrum can be used temporarily by other systems when the primary system is idle. This concept creates the so-called secondary spectrum markets.

2.6 Summary

In this chapter, we provided an overview of AAA signaling protocols including RADIUS and Diameter and demonstrated the range of applications in wireless networks within all tiers: the radio, core, and service tiers. We discussed how AAA protocols are used to authenticate and authorize users' access and to report accounting for service usage. We also explained the relationship between the AAA system and the Business Support System (BSS) for prepaid and postpaid AAA customers. We then provided examples on the AAA signaling and operation within all tiers. We concluded the section by discussing further AAA research trends beyond security including system planning and scalability, signaling delay reduction, accounting signaling optimization, and AAA design for emerging networks such as adhoc and L2 optical networks. In the next chapter, we address the fundamental challenge of AAA system planning and derive analytical models for fixed and mobile networks in centralized and distributed deployments.

Chapter 3 AAA System Planning Models

3.1 Introduction

Due to the traditional separation of radio and IP engineering in 3G cellular networks, large operators plan the capacity of their AAA systems by over-provisioning. This is mainly because under provisioned systems would result in blocking users from access or dropping accounting messages, leading to loss of revenue. Although the growth of the AAA signaling is imminent, which is expected to turn over-provisioning inefficient and hard to scale, alternative design guidelines are currently missing.

The AAA system design depends on the knowledge of the AAA signaling rate. From an analytical perspective, the evaluation of the AAA signaling load, however, is nontrivial and can not be simply evaluated by direct application of tele-traffic models used in the data plane. This is due to the need to consider the system configuration (i.e., fixed or mobile), protocol aspects, mobility, and session durations. The interaction of these attributes turns the AAA system planning problem which we address in this chapter unique and interesting.

In this chapter, we derive AAA system planning models under generic deployment assumptions. We start our development by investigating AAA signaling in fixed deployments and then extend the derived model to mobile deployments. In the consideration for mobile networks, we first consider centralized AAA system configurations and then generalize the analysis to distributed AAA systems including roaming scenarios. The developed framework intertwines concepts from renewal theory, stochastic processes, transient Markov chains, and complex variable analysis. Since mobility plays a key role in the developed planning models, we dedicate Section 3.8 for discussing our novel contributions towards the development of a generalized handoff modeling framework based on general session duration and residence time statistics, mobility patterns, and user concentrations.

3.2 Supporting Publications

1. W. Bziuk, S. Zaghloul and A. Jukan, "A New Framework for Characterizing the Number of Handoffs in Cellular Networks," European Transactions on Telecommunications, 20(7), Sep 2009, pp. 689-700
2. S. Zaghloul, W. Bziuk and A. Jukan, "A Novel Analytical Framework for Mobility Modeling in All-IP Wireless Systems," proceedings of the 21st International Teletraffic Congress (ITC 21) proceedings, Paris, France, Sep 2009
3. S. Zaghloul, A. Jukan, "Signaling Rate and Performance for Authentication, Authorization, and Accounting (AAA) Systems in all-IP Cellular Networks," IEEE Transactions on Wireless Communications, 8(6), pp. 2960-2971 Jun 2009
4. W. Bziuk, S. Zaghloul and A. Jukan, "The Spatial Effect of Mobility on the Mean Number of Handoffs: A New Theoretical Result," proceedings of the IEEE International Communications Conference (ICC'09), Dresden, Germany, Jun 2009
5. S. Zaghloul, W. Bziuk, A. Jukan, "Signaling and Handoff Rates at the Policy Control Function (PCF) in IP Multimedia Subsystem (IMS)," IEEE Communications Letters Journal, pp. 526-528, Jul 2008
6. S. Zaghloul, A. Jukan, "On the Performance of the AAA Systems in 3G Cellular Networks," proceedings of the IEEE International Communications Conference (ICC'07), Glasgow, United Kingdom, pp. 2103-2108, Jun 2007
7. S. Zaghloul, A. Jukan, "Relating the AAA and the Radio Access Rates in 3G Cellular Networks," IEEE Communications Letters Journal, 11(4), pp. 363-365, Apr 2007
8. W. Bziuk, S. Zaghloul, A. Jukan, "A Novel Framework for Handoff Analysis Under Generalized Session and Mobility Statistics," under conference review
9. S. Zaghloul, A. Jukan, "A Generic Framework for Planning AAA Signaling in Centralized and Distributed Network Deployments," under journal submission

3.3 General Chapter Assumptions

1. User session requests, Ψ, arrive following a Poissonian process with mean arrival rate, λ req/s.
2. The session duration, S, is assumed to be negative exponentially distributed with a mean of $E[S] = E_s$. In Section 3.8, we relax this assumption when calculating the mean number of handoffs.
3. No blocking occurs at the Network Access Server (NAS) for the given session arrival rate from the RAN. This is realistic due to the high capacity of the network access servers which support thousands of simultaneous sessions [90].

4. AAA packets are not fragmented and are of fixed length. This is the case in current networks since most user profile sizes are approximately equal.

5. Without loss of generality, reauthentications are always successful for already authenticated users.

3.4 AAA Signaling Rate in Fixed Environments

This section derives an analytical model for the mean and the variance of the AAA signaling rate sent by a large NAS (e.g., see [90]) to an AAA server given the user access rate from the RAN. In our analysis, we also study the effect of retransmissions due packet losses in the network by assuming the use of the RADIUS protocol [40, 41] which performs worse than its successor protocol, the Diameter protocol [45] in terms of the number of retransmissions[1] and hence provides a upper bound on the effect of retransmissions on the protocol performance.

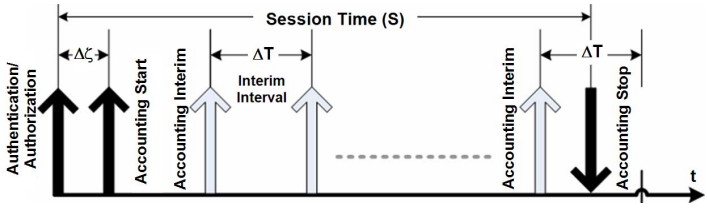


Figure 3.1: RADIUS protocol messages for a typical user session (adapted from [91, 92]).

As illustrated in Fig. 3.1, the outbound RADIUS traffic (from the NAS) is comprised of two basic message types: *Access Requests* (also called *Authentication/Authorization Requests*) and *Accounting Requests*. If not lost, every *Access-Request* results in either an *Accept* or a *Reject* response. If the user is accepted, *Accounting* messages are generated, including a *Start, Interim(s)* and a *Stop*. Every message is acknowledged by an *Accounting-Ack*. Reliability is attained by requiring a response for each message which times out if a response is not received within a predefined timeout period (TO). It is then up to the NAS to either retransmit to the same AAA server, another server, or even drop the request. The maximum number of retransmissions (N) is a NAS configuration parameter (see [93]). Unlike Access Requests, accounting requests are sent regardless of the acknowledgement of the previous ones (e.g., *Stops* are sent regardless of the reception of *Interims*).

[1] Recall that RADIUS uses the intrinsically unreliable User Datagram Protocol (UDP) protocol while Diameter uses reliable mechanisms such as Transport Control Protocol (TCP) or Stream Control Transmission Protocol (SCTP) for transport.

3.4.1 Assumptions for AAA System Planning in Fixed Environments

1. The maximum number of request retransmissions is N.
2. The Packet Error Rate (PER) for the link between the NAS and the AAA server is denoted as p where we typically have $p < 1\%$.
3. The Time out (TO) period is much shorter than the *Accounting-Interim-Interval* denoted as Δ_T.

3.4.2 Mathematical Model

This section starts by characterizing the mean AAA traffic rate (ξ) generated in response to a Poisson distributed subscriber RAN access process, denoted as $\Psi(\lambda, t)$. Let ξ_A and ξ_{Re} denote the authentication and reauthentications rates respectively. Let ξ_{Start}, ξ_{Int}, and ξ_{Stop} denote the outbound accounting *Start*, *Interims*, and *Stop* rates respectively. Then, the mean AAA rate from all NASes, $E[\xi]$, can be written as the sum of its message components as,

$$E[\xi] = E[\xi_A] + E[\xi_{Re}] + E[\xi_{Start}] + E[\xi_{Int}] + E[\xi_{Stop}] \tag{3.1}$$

Assuming independent packet losses over the link between the NAS and the AAA server and assuming a fixed PER of p, then the conditional probability density function of the number of individual packet transmissions, denoted as $f_x(k/p)$ is,

$$f_x(k/p) = \frac{(1-p)p^{k-1}}{1-p^{N+1}} \text{ where } k = 1, \cdots, N+1 \tag{3.2}$$

The Probability Generating Function (PGF) of the number of request transmissions, $f_x(k/p)$, is then given as,

$$X(z/p) = \sum_{k=1}^{N+1} \frac{(1-p)p^{k-1}}{1-p^{N+1}} z^k = \frac{1-p}{1-p^{N+1}} \frac{1-(pz)^{N+1}}{1-pz} \tag{3.3}$$

Using the first and the second derivatives of the PGF in (3.3) and with algebraic manipulations, it can be shown that the mean and the variance of the number of request transmissions are,

$$E[X/p] = 1 + \frac{p}{1-p} - \frac{(N+1)p^{N+1}}{1-p^{N+1}} \tag{3.4}$$

$$Var[X/p] = \frac{p + p^{3+2N} - p^{N+1}[(1+N)^2(1+p^2) - 2N(2+N)p]}{[(1-p)(1-p^{N+1})]^2} \tag{3.5}$$

Since the sizes of accounting acknowledgements are often much smaller than accounting requests [41, 45], it is assumed that they are never lost. However, since *Access-Accept* messages include subscriber authorization data, the sizes of such messages are comparable to *Access-Requests*, and thus their loss probability (q) is,

$$q = 1 - (1-p)^2 = 2p - p^2 \tag{3.6}$$

Since the RAN requests Ψ follow a Poisson distribution and since the number of packet transmissions follows the $f_x(k/p)$ distribution, the PGF of the authentication arrival process is $\xi_A[z] = \Psi[X[z/q]]$. The access response (i.e.,accept/reject) reception probability given a maximum of N retransmissions can be expressed as,

$$\delta = \sum_{k=1}^{N+1}(1-q)q^{k-1} = 1 - q^{N+1} = 1 - (2p-p^2)^{N+1} \tag{3.7}$$

If the *Access-Accept* rate (i.e., percentage of successful authentications) is fixed and denoted as p_a, the successful authentications process (Ω) may be approximated by splitting the driving RAN Poisson process $\Psi(\lambda)$ by δp_a as, $\Omega = \Psi(\delta p_a \lambda)$. Here the Poisson approximation is justified as the error caused by retransmissions and rejections is negligible with a low p and a low access rejection rate. At steady state (i.e., after $t >> E_s$), the first and second order statistics of ξ_{Stop} and ξ_{Start} are approximately equal. Considering retransmissions, the PGFs of the outbound accounting *Start* and *Stop* are,

$$\xi_{Start}[z] = \xi_{Stop}[z] = \Omega[X[z/p]] \tag{3.8}$$

The steady state rate for the *Interims* is obtained by noting that during a short rate measurement period (i.e., of the order of TO), the *Interims* will mostly correspond to different user sessions, since $TO << \Delta_T$. Typically, TO is in the order of seconds while Δ_T is in the order of minutes. The *Interims* arrival process is characterized by summing the *Interims* from all active sessions as,

$$\begin{aligned} \xi_{Int}[z] &= \Omega_I[z] = \sum_{j>0} P(S > j\Delta_T)\Omega[X[z/p]] \\ &= \sum_{j>0} e^{-\frac{j\Delta_T}{E_s}}\Omega[X[z/p]] = \frac{\Omega[X[z/p]]}{e^{\frac{\Delta_T}{E_s}} - 1} = \varphi\Omega[X[z/p]] \end{aligned} \tag{3.9}$$

Note that $\varphi = \left[e^{\frac{\Delta_T}{E_s}} - 1\right]^{-1}$ is the mean number of *Interims* for an exponentially distributed session time calculated as the mean of $floor[S/\Delta_T]$. This is clear by comparing the survivor definition of the expectation in [94] to (3.9). Finally, the reauthentications signaling rate is calculated similarly to the accounting interims rate as by substituting the interim interval Δ_T with the *Authorization Lifetime*, Δ_M as,

$$\xi_{Re}[z] = \xi_{Int}[z] \mid_{\Delta_M = \Delta_T} = \varphi(\Delta_M)\,\Omega[X[z/p]] \tag{3.10}$$

3.4.2.1 The AAA Signaling Rate Model in Fixed Networks

The mean AAA signaling rate, $E[\xi]$, is obtained by substituting results from (3.4)-(3.9) into (3.1) and rearranging, as,

$$E[\xi] = \lambda(E[X/q] + \delta p_a[\varphi(\Delta_T) + \varphi(\Delta_M) + 2]E[X/p]) \tag{3.11}$$

For the variance, we are interested in the variance of the AAA signaling load within a short measurement period. Recall that the variance for the sum of a K correlated random variables Y_i are given as,

$$Z = \sum_{i=1}^{i=K} Y_i \quad , \quad \mathrm{Var}(Z) = \sum_{i=1}^{i=K} \mathrm{Var}(Y_i) + \sum_{i=1}^{i=K} \sum_{\substack{j=1 \\ i \neq j}}^{j=K} \mathrm{Cov}(Y_i, Y_j) \tag{3.12}$$

In general, we know that unlike the mean value, the variance depends on the correlation properties between the random components of the signaling messages. Hence, the variance of the AAA signaling rate, $Var[\xi]$, is expressed as the sum of the variance due to each component plus the covariance ϑ between them as,

$$Var[\xi] = Var[\xi_A] + Var[\xi_{Start}] + Var[\xi_{Int}] + Var[\xi_{Re}] + Var[\xi_{Stop}] + \vartheta \tag{3.13}$$

The covariance term ϑ in (3.13) can be obtained by observing the correlation between the different components of the AAA signaling messages during a short measurement period of TO such that ($\lambda >> TO^{-1}$) for the same user session as follows,

1. There is a high correlation between the rates of the *Access-Accept* messages and the first transmission of *Accounting-Start* messages given as (1-p)Ω. This is expected as these messages almost correspond to the same event. Here, the correlation coefficient is approximated as $(TO^{-1}[TO - \Delta\xi] \approx 1)$ for low delay networks (i.e., when $TO >> \Delta\zeta$). Thus, this component contributes with a $2Var[\Omega]\sqrt{1-p} \times 1$ to the covariance.

2. Interims and reauthentications may occur together depending on the settings of the reauthorization and interim intervals, Δ_M and Δ_T. If $\Delta_M = \Delta_T$, then we have practically the same event and hence the contribution to the covariance term is $2\mathrm{Var}(\Omega)(1-p)\varphi(\Delta_T)$. Otherwise, we need to consider this event depending on how often these two periods coincide. In other words, for the interims and reauthentication traffic, there will be an independent term that does not contribute to the covariance when both periods do not coincide and another that will be considered when both periods coincide within the measurement period characterized by the least common multiple (lcm) of both periods denoted as Δ_G. For example, if $\Delta_M = 2\Delta_T$ then the lcm is Δ_M and hence the contribution to the covariance term is $2\mathrm{Var}(\Omega)(1-p)\varphi(\Delta_G)$.

3. The dependencies among the other AAA message types within a TO period are minimal. In other words, having very short sessions with authentication requests, starts, and stops within a very short time period is highly unlikely. Furthermore, since *Interims* and *Stops* are correlated only if the session stops right after sending an interim message, this event is also unlikely and can be neglected. Finally, since the interim interval, Δ_T, is usually high, then interims can not occur along with authentications and accounting starts within a short measurement window.

Hence, the covariance term is given as $\vartheta = 2\text{Var}(\Omega)\sqrt{1-p} + 2\text{Var}(\Omega)(1-p)\varphi(\Delta_G)$, where $\Delta_G = lcm\{\Delta_T, \Delta_M\}$. Therefore, substituting (3.3)-(3.9) into (3.13), the variance of the AAA signaling load, $\text{Var}(\xi)$, is given as,

$$\begin{aligned} Var[\xi] &= Var[\Psi]E[X^2/q] + Var[\Omega]([m+2]E[X^2/p] + 2\sqrt{1-p} + 2(1-p)\varphi(\Delta_G)) \\ &= \lambda E[X^2/q] + \lambda\delta p_a([m+2]E[X^2/p] + 2\sqrt{1-p} + 2(1-p)\varphi(\Delta_G)) \end{aligned} \quad (3.14)$$

Fig.3.2 illustrates the mean AAA signaling rate and its variance as functions of the ratio of the accounting interim interval to the mean session duration. The figure confirms our intuition that the AAA signaling rate decays and approaches an asymptote as the normalized interim interval increases (i.e., when no *Interims* are sent). However, since the main purpose of the *Interims* is to protect against revenue losses in case of hardware/network failures, cellular operators cannot avoid their deployment; therefore necessitating prudent assessment of the tradeoff between increased accounting reliability (i.e., with large number of *Interims*) and the cost of injecting more *Interims* into the network. Although operator network congestion is unlikely in many cases, the AAA server resources wasted to process a large volume of Interims instead of handling other requests can be an issue. The same analogy applies to the upstream billing systems that correlate accounting records to produce the final usage bill.

Another conclusion that we draw from our model is that while retransmissions improve the system's reliability, raising the maximum number of retransmissions (N) barely affects the resulting signaling rate. This is due to the low packet errors in today's wired networks, typically below 1%. This is illustrated in Table 3.1. The results in Table 3.1 indicate that even for a relatively high PER and with aggressive retransmissions from the RADIUS protocol, which is unlike the more conservative Diameter protocol, retransmissions do not play a significant role in the analysis of the signaling load and can be safely ignored.

We now proceed by relaxing the exponential session assumption and thus obtaining a more general formula for φ in (3.9).

Table 3.1: RADIUS rate statistics for various number of retransmissions [92], N, ($p = 1\%, m = 2$).

Session arrival rate/s	Mean(E[ξ])			Variance(Var[ξ])		
	N=1	N=2	N=10	N=1	N=2	N=10
10	51	51	51	71.7	71.7	71.7
130	662.5	663.2	663.2	931.4	932.6	932.6

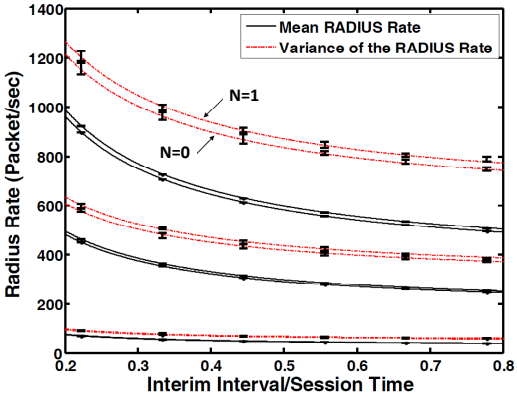


Figure 3.2: Mean and variance of the outbound RADIUS rate in fixed networks as function of the normalized *Interim-Interval* for different arrival rates (adapted from [92]). [Simulation parameters: confidence level 95%, p=1%, 5 iter/run, run =6 hrs, per second measurements recorded]

3.4.2.2 Relaxing the Exponential Session Duration Assumption

In our earlier analysis in (3.9), we calculated the mean number of interims during the session by assuming an exponentially distributed session duration. Now, we show how to obtain the number of interim and re-authentication messages during any arbitrary duration H which does not necessarily follow the exponential distribution. Using the Complementary Commutative Distribution Function (CCDF) of H, it can be shown (see Appendix A.1.1) that the mean number of interims per session is given as,

$$\varphi(\Delta_T) = E[\lfloor \frac{H}{\Delta_T} \rfloor] = \sum_{j=1}^{\infty} \bar{F}_H(j\Delta_T) \tag{3.15}$$

To get an insight to the general formula in (3.15), let us reconsider our exemplary case of a single service with an exponentially distributed session duration (i.e., $\bar{F}_H(h) = e^{-\frac{h}{E_S}}$).

It directly follows that (3.15) simplifies to,

$$\varphi(\Delta_T) = \sum_{j=1}^{\infty} e^{-\frac{j\Delta_T}{E_s}} = \frac{1}{e^{\frac{\Delta_T}{E_s}} - 1} \tag{3.16}$$

Clearly the result in (3.16) matches that in (3.9). A useful result that we later use is when H follows the log normal distribution. The log normal distribution is particularly interesting as it is widely used to fit measurements for voice call and data session durations [95–99]. Since the complementary distribution for the LogNormal is $\bar{F}_H(h) = \frac{1}{2}\text{erfc}\left(\frac{\ln h - \mu}{\sqrt{2\sigma^2}}\right)$, then using (3.15), it follows that,

$$\varphi(\Delta_T) = \frac{1}{2}\sum_{k=1}^{\infty} \text{erfc}\left(\frac{\ln(k\Delta_T) - \mu}{\sqrt{2\sigma^2}}\right) \tag{3.17}$$

where the parameters μ and σ are given in terms of the mean duration E_H and its coefficient of variation C_H, which denotes the ratio of the standard deviation to the mean of H, as,

$$\mu = \ln(E_H) - \frac{(\sigma)^2}{2} \quad , \quad \sigma^2 = \ln\left((\mathrm{C}_H)^2 + 1\right) \tag{3.18}$$

Now that we have characterized the AAA signaling rate in fixed environments, we extend the result in (3.11) to include the effect of mobility in Section 3.5.

3.4.3 AAA Fixed Network Model's Limitations

1. *The session arrival process*: The accuracy of the model's predictions for the variance of the AAA signaling load depends on the accuracy of the Poissonian assumption for the session arrivals. While this is justified for the current session types which are primarily generated by many users, this may not be accurate for future broadband wireless networks supporting rich sessions with mixes of voice and data flows.

2. *The mobility of the users*: The signaling model applies to fixed network environments (e.g., fixed WiMAX deployments or nomadic mobility which refers to very low mobility profiles like users using their laptops). The generalization for arbitrary mobility profiles is the subject of Section 3.5.

3. *The processing power of the AAA system*: Processing delays of the AAA messages are deliberately ignored in our analysis as they vary considerably depending on the AAA implementation. They can be easily incorporated into our analysis by multiplying each message type by a cost. Our model also assumes that neither the NAS nor the AAA system will alter the session arrival process by introducing blocking or jitter.

4. *The users' quota*: We assume a postpaid model where users are billed offline hence the users' quota does not play a role in our model (see Section 2.3). Forced session terminations can happen due to the exhaustion of the users' quota. This aspect is not considered in our model and the likelihood of forced termination due to quota exhaustion is currently an open question. Aspects of this problem were addressed in [84–86] which evaluate the likelihood of session dropping due to quota depletion in a time-based system offering a single service.

3.5 Signaling in Mobile Environments: Basic Model

In this section, we study the practical case in mobile networks, where the users' movement in the network results in AAA signaling. Since the durations the users spend in a NAS region may vary depending on the mobility profile of the users (i.e., fast moving or slow), the time duration the NAS serves a user no longer equals the session duration and hence needs further investigation. In the context of this section, the NAS functional entity refers to the AGW. The AAA (especially accounting) signaling can be triggered on a per sector basis or whenever the users move between regions covered by different AGWs. Both cases can be addressed in almost the same manner and hence we focus on the movement between AGW region for simplicity. Since as we previously mentioned in Chapter 3, both RADIUS and Diameter protocols largely incorporate the same AAA signaling message types and protocol procedures, we adopt the message names from Diameter [40, 41, 45, 61, 100]. In the following subsections, we generalize the fixed network model in 3.4 to account for users' mobility in the network.

3.5.1 Reference Architecture

As we mentioned previously in Chapter 3, in all-IP cellular architectures [42, 101], AGWs serve multiple base station areas[2] as shown in Fig. 3.3. AAA signaling is triggered as users initiate or terminate their sessions, and when they move between AGW regions.

Let us now investigate the AAA signaling pertaining to users' sessions as they move from an AGW region into another. As shown in Fig. 3.4, when a user initiates a mobile session, the radio access network triggers AAA signaling at the corresponding AGW towards the AAA system. When a session is established (step 1), Diameter authentication exchanges (i.e., AA-Mobile-Node-Request, AMR) are conducted with the AAA system to authenticate and/or authorize the incoming session. The authentication response (i.e., AA-Mobile-Node-Answer, AMA) carries the user's profile and network settings back

[2] An AGW is a generic term used to refer to any first IP gateway; examples of AGWs are ASN-GW in WiMAX, PDSN in the Third Generation Partnership Project II (3GPP2) networks, or S-GW in the Third Generation Partnership Project (3GPP) Rel6+ systems.

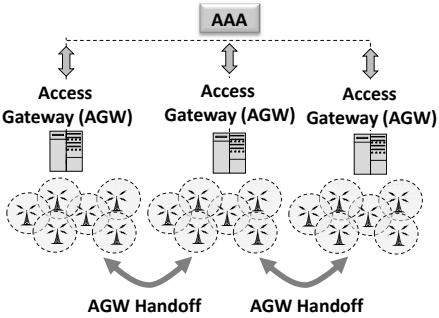


Figure 3.3: A simplified "all-IP" system (adapted from [102]).

to the requesting gateway. One of these settings is the Authorization-Lifetime attribute which is used to indicate the time by which the mobile node must re-authenticate once it expires. In our example, a re-authentication takes place in step 4. Upon successful authentication, an Accounting Request message, ACR type Start, is sent (step 2). The AAA acknowledges the receipt of the ACR message by sending an accounting answer message (ACA). The accounting ACR Start message is typically followed by periodic ACR type (Interim) messages reporting the latest subscriber's usage every Acct-Interim-Interval (steps 3, 5) [45]. As we mentioned in Chapter 2, accounting interim messages are used to periodically meter users' sessions and thus minimize revenue losses should the network suffer from unexpected failures [6, 45]. When a handoff occurs between AGW 1 and 2, the accounting session at the source AGW is terminated with an ACR type Stop message (step 6), while a new accounting session is sent by the target AGW after optionally authenticating the user (step 7). Steps similar to (1-6) occur at the new AGW. Once the session is terminated (step 14), an ACR type (Stop) message is sent reporting the final subscriber's usage.

3.5.2 Assumptions

In addition to the assumptions in Section 3.3, we have the following assumptions,

1. The AGW residence times, R, are independent and identically distributed following the Gamma distribution with a mean of $E_r = k_r\theta_r$. k_r and θ_r are the shape and scale parameters, respectively. Here, the shape parameter is the reciprocal of the square of the coefficient of variation while the scale parameter tells how large the distribution is spread-out. Note that we choose the Gamma distribution for the AGW residence time (similar to many articles in the literature such as

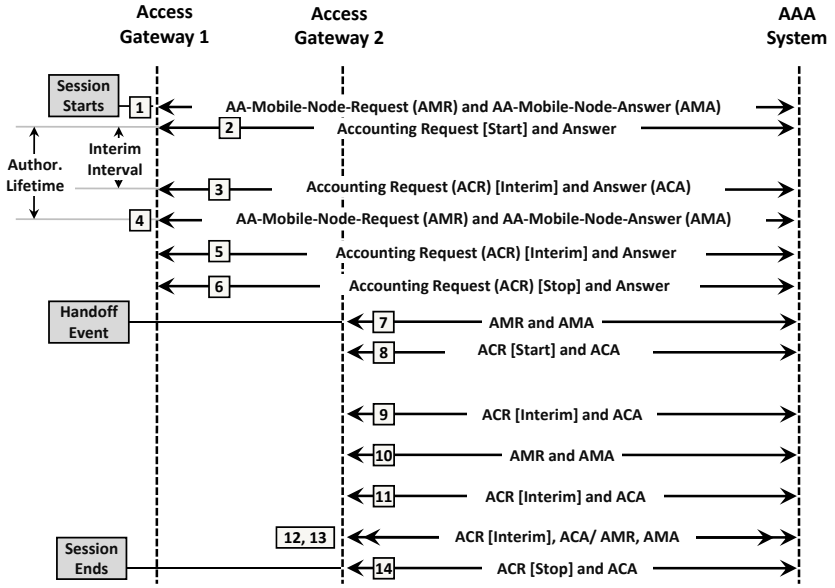


Figure 3.4: Typical Diameter signaling messages (adapted from [102]).

[103]) as it is known to offer a good approximation for the lognormal distribution [104], the widely encountered distribution for cell residence times from field measurements [17].

2. For simplicity, we assume no retransmissions for AAA requests.

3.5.3 Mathematical Model

3.5.3.1 Mobility Description

In this section, we extend the result in (3.11) for the signaling rate in fixed environments to include the effect of mobility. To do so, let us first define two important random variables that we will extensively use in the analysis: the AGW residence time, and the AGW holding time. Our definitions below are similar to the channel holding time and the cell residence time definitions which are widely used in foundational research studies in the area of cellular systems performance [17, 18, 37].

- *AGW residence time, R*: This quantity refers to the time the user spends in an access gateway area irrespective of the session's activity [17, 18, 37]. When a session starts in an AGW region, the residual of the residence time, $\tilde{R}$, is used to refer to the time the user spends in the initial AGW before departure. The use of the residual of the residence time is needed as the users' movement within an AGW coverage area and the session start are two independent events. However, the residence time R is used for all subsequent handoffs. Due to the large AGW coverage area, only users nearby the border regions leave the AGW region. Thus, it is expected that the AGW residence time have a large coefficient of variation. Thus, methods of clustering or profiling should be used to catagorize users according to their residence times (e.g., low mobility users with large mean residence time and high mobility users with low residence time). In the rest of the discussion, the AGW residence time refers to the residence time of the profile under consideration.

- *AGW holding time, H*: This quantity only applies to active sessions and is defined as the time from a session start or from the most recent handoff event until the next handoff event or the session termination. In other words, it is the minimum of the remaining lifetime of the session duration and the time a user spends within the AGW region (i.e., the AGW residence time) [37].

Without loss of generality, the residence time in our analysis is assumed to be Gamma distributed. Other distributions such as Hyper Erlang [17] can also be accommodated in the same manner. However, fitting such general distributions to measured data of the residence time requires the use of more sophisticated schemes such as the Expectation Maximization (EM) schemes as in [105]. Once fitted, our results for the gamma distribution can be readily used. This is because the Hyper Erlang is simply a scaled sum of Erlangian/Gamma terms. The Gamma PDF, Commutative Distribution Function (CDF), and CCDF are given as,

$$f_R(x) = \frac{x^{k_r-1}e^{-\frac{x}{\theta_r}}}{\theta_r^{k_r}\Gamma(k_r)} \quad , \quad F_R(x) = \frac{\gamma\left(k_r, \frac{x}{\theta_r}\right)}{\Gamma(k_r)} \quad , \quad \bar{F}_R(x) = \frac{\Gamma\left(k_r, \frac{x}{\theta_r}\right)}{\Gamma(k_r)}$$

where $\gamma(k,x)$ and $\Gamma(k)$ are the incomplete and complete gamma functions as defined in Appendix B.2. The shape parameter k_r is obtained by taking the reciprocal of the coefficient of variation of the measured quantity (i.e., the residence time measurements here) while the scale parameter is obtained by dividing the mean measurement (i.e, the expected residence time, $E[R] = E_r$) by the calculated shape parameter as,

$$\begin{aligned} k_r &= \frac{1}{C_r^2} = \left(\frac{\text{Mean of the residence time measurements, } E_r}{\text{Standard deviation of the residence time measurements}}\right)^2 \\ \theta_r &= (\text{Mean of the residence time measurements, } E_r) \times 1/k_r \end{aligned} \tag{3.19}$$

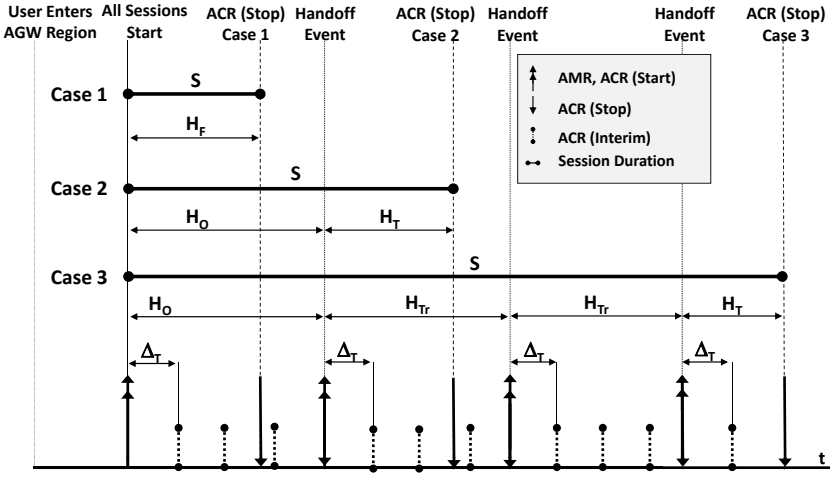


Figure 3.5: AAA traffic model [102] [*Reauthentications* (AMR) are omitted for clarity].

The residual of the residence time, $\tilde{R}$, is given by using the definition of the residual distribution as [94],

$$f_{\tilde{R}}(\tilde{r}) = \frac{\bar{F}_R(\tilde{r})}{E_r} = \frac{\Gamma\left(k_r, \frac{\tilde{r}}{\theta_r}\right)}{E_r \Gamma(k_r)} \tag{3.20}$$

We now turn our attention to how the residence and holding time concepts play key roles in the behavior of the session in mobile networks. As shown in Fig. 3.5, any user's session, S, falls under one of the following three categories with respect to the number of handoffs it makes during its lifetime:

Case (1) *No handoffs:* This case occurs when the session duration is shorter than the remaining time for the user to leave the AGW region (i.e., residual residence time). This is commonly assumed in the literature [17, 37] as the moments when users emerge into an area and when they initiate sessions are not necessarily aligned.

Case (2) *Only one handoff:* Occurs if the users make exactly one AGW handoff during their session. Hence, the session will be served by exactly two AGWs.

Case (3) *Multiple handoffs:* This case occurs if the user makes at least one AGW handoff during her session. This case generalizes the previous one to include *transit* AGWs.

Notice that we consider such cases for analytical purposes in order to obtain the distribution of the AGW holding time which we later use extensively in our analysis. The reason why we need to evaluate such complex distributions is due to the fact that each AGW generates AAA signaling only for the duration it serves the session. Therefore, we need to formally study the characteristics of such AGW holding times in order to evaluate the mean number of reauthentications and interims similar to the procedure in Section 3.4 in (3.15). In this context and as shown in Fig. 3.5, we see four general types of AGW holding times which we denote as $H_i, i \in \{F, O, Tr, T\}$ as follows:

- *Full Sessions,* H_F: This type of AGW holding times matches the full session duration, S, and occurs when no AGW handoffs take place.
- *Originating Sessions,* H_O: This type of AGW holding times refers to the duration that sessions spend in the *initial* AGW in which they start. In this case, the session incurs at least one handoff.
- *Transit Sessions,* H_{Tr}: This type of AGW holding times refers to the duration that passing (i.e., transiting) sessions spend within an AGW region. Transit sessions neither initiate nor terminate in the AGW under consideration. Transit sessions occur when more than one handoffs takes place during the sessions' life times.
- *Terminating Sessions,* H_T: This type of AGW holding times refers to the duration that sessions spend in the *last* AGW where they terminate. This category occurs when at least one handoff takes place which makes it different from full sessions.

Let us now analyze the conditions for each holding time to occur.

- *Full Sessions,* H_F: If there are no handoffs in the session (i.e., case 1), the AGW holding time (i.e., H_F) is the conditional duration of the user's session being smaller or equal to the residual residence time (i.e., $S \leq \tilde{R}$). Recall that the residence time incurred in the first AGW is always characterized by the residual AGW residence time, $\tilde{R}$.
- *Originating Sessions,* H_O: For H_O to occur, the conditional duration of the residual residence time should be less than or equal to session time (i.e., $\tilde{R} \leq S$). This happens when the session incurs at least one handoff.
- *Transit Sessions,* H_{Tr}: This holding time happens if the remainder[3] of the session lifetime, S_R^+, is greater than the AGW residence time R as $S_R^+ > R$. Because of the memoryless property of the exponential distribution, the remainder of the session time statistically equals the session duration and hence H_{Tr} happens if

[3]Here, we do not refer to the residual session duration but rather to the conditional random variable $(S_R^+ = S - H_O - H_{Tr} \ldots \mid S > 0)$. The residual of the session time $\tilde{S}$ is used to approximate S_R^+ as in [37].

$S > R$. Since this type only happens if the session incurs more than one AGW handoff, the number of the H_{Tr} periods is one less than the number of handoffs, K (i.e., $K-1$).

- *Terminating Sessions, H_T*: For H_T to occur the conditional duration of the remainder of the session time, S_R^+, should be less than the AGW residence time R as $S_R^+ < R$. Due to the memoryless property, the condition becomes $S < R$. This happens when the session incurs at least one handoff.

3.5.3.2 AAA Signaling Analysis

From (3.1) recall that the AAA traffic consists of authentication and accounting traffic. Since we now consider a network with N_{AGWs} with no retransmissions, the mean authentication rate, $E[\xi_A]$, can be rewritten as the sum of authentication requests from new sessions plus authentication requests due to handoffs that a session makes in possible AGWs in the network.

$$E[\xi_A] = \sum_{i=1}^{N_{\text{AGWs}}} (E[K]+1)\,\lambda^{(i)} \tag{3.21}$$

where $\lambda^{(i)}$ denotes session arrivals from the i^{th} AGW in the network.

Since there is only one message of the type (authentications, accounting starts and stops) in a given session, it follows that at steady state, the mean rate of these types is approximately equal. This is because for operational networks the authentication success rate is around unity (i.e., $p_a \approx 1$).Thus, the rates of these messages are,

$$E[\xi_{Start}] = E[\xi_{Stop}] = p_a E[\xi_A] \tag{3.22}$$

At this point, we turn our attention to the evaluation of the interim and reauthentication rates using the definitions of the holding times. To do so, we first evaluate the distributions of the AGW holding times, H_i, $i \in \{F, O, Tr, T\}$ as in Fig. 3.5 and use the results to find the number of interims and reauthentications.

3.5.3.3 Evaluating the Holding Times

In this section, we apply results in (A.7) from Appendix A.1.2 to evaluate the AGW holding times, $H_F \ldots H_T$ as follows,

- *The CDF of H_F*: This quantity is defined as $Pr\left(S \le h \mid S \le \tilde{R}\right)$ as is found by integrating the joint probability of S and $\tilde{R}$ in the region limited by h and by

dividing the result by the probability that $(S \leq \tilde{R})$ as,

$$F_{H_F}(h) = Pr\left(S \leq h \mid S \leq \tilde{R}\right) = \frac{\int_0^{\infty} \int_0^{min(y,h)} f_S(x)\, dx f_{\tilde{R}}(y)\, dy}{Pr\left(S \leq \tilde{R}\right)} \tag{3.23}$$

Using (A.6), the numerator (3.23) in is evaluated as,

$$\int_{y=0}^{\infty} \int_{x=0}^{min(y,h)} f_S(x)\, f_{\tilde{R}}(y)\, dxdy = \int_0^h F_S(y)\, f_{\tilde{R}}(y)\, dy + F_S(h)\, \bar{F}_{\tilde{R}}(h)$$

The denominator $Pr\left(S \leq \tilde{R}\right)$ represents the probability of making no handoffs, p_0, and is given as $1 + \left(\left(\frac{\theta_h}{\theta_r}\right)^{k_r} - 1\right)\frac{E_s}{E_r}$ as will be shown later in (3.35). Using (B.6) and (B.7) from Appendix B.2, then integrating by parts it can be shown that $F_{H_F}(h)$ can be written as,

$$F_{H_F}(h) = \frac{B_0 - E_s\left(\frac{\theta_h}{\theta_r}\right)^{k_r}\Gamma\left(k_r, \frac{h}{\theta_h}\right)}{B_0} + e^{\frac{-h}{E_s}}\frac{(h+E_s)\,\Gamma\left(k_r, \frac{h}{\theta_r}\right) - \theta_r\Gamma\left(k_r+1, \frac{h}{\theta_r}\right)}{B_0}$$

$$B_0 = \left[E_r + E_s\left(\left(\frac{\theta_h}{\theta_r}\right)^{k_r} - 1\right)\right]\Gamma(k_r) \tag{3.24}$$

where we defined the parameter θ_h as,

$$\theta_h = \frac{\theta_r E_s}{E_s + \theta_r} \tag{3.25}$$

- *The CDF of H_O*: This quantity is given as $Pr\left(S \leq h \mid \tilde{R} \leq S\right)$ as is found by integrating the joint probability of S and $\tilde{R}$ in the region limited by h and dividing the result by the probability that $(\tilde{R} \leq S)$ given as $Pr\left(\tilde{R} \leq S\right) = \int_0^{\infty} \int_0^{y} f_{\tilde{R}}(x)\, dx f_S(y)\, dy$. Following similar analysis as in (3.24), it can be shown that $F_{H_O}(h)$ is given as,

$$F_{H_O}(h) = Pr\left(S \leq h \mid \tilde{R} \leq S\right) = \frac{\int_0^{\infty} \int_0^{min(y,h)} f_{\tilde{R}}(x)\, dx f_S(y)\, dy}{Pr\left(\tilde{R} \leq S\right)} \tag{3.26}$$

$$= \frac{\left(\frac{\theta_h}{\theta_r}\right)^{k_r}\gamma\left(k_r, \frac{h}{\theta_h}\right)}{\Gamma(k_r)\left(\left(\frac{\theta_h}{\theta_r}\right)^{k_r} - 1\right)} - \Gamma(k_r) + e^{\frac{-h}{E_s}}\Gamma\left(k_r, \frac{h}{\theta_r}\right)$$

- *The CDF of H_{Tr}*: This quantity is given as $F_{H_{Tr}}(h)$=$Pr\left(S_R^+ \leq h \mid R \leq S_R^+\right)$. Due to the memoryless property of the exponential distribution $S_R^+ \equiv S$. Therefore $F_{H_{Tr}}(h)$ is expressed as the integration of the joint probability of S and R in the

region limited by h and dividing the result by the probability that $(R \leq S)$ given as $Pr(R \leq S) = \int_0^\infty \int_0^y f_R(x)\,dx f_S(y)\,dy$. Following similar analysis as in (3.24), it can be shown that $F_{H_{Tr}}(h)$ follows the Gamma distribution with with shape and scale parameters of k_r and θ_h as,

$$F_{H_{Tr}}(h) = Pr(S \leq h \mid R \leq S) = \frac{\int_0^\infty \int_0^{min(y,h)} f_R(x)\,dx f_S(y)\,dy}{Pr(R \leq S)}$$
$$= \frac{\gamma\left(k_r, \frac{h}{\theta_h}\right)}{\Gamma(k_r)}\,,\ \theta_h = \frac{\theta_r E_s}{E_s + \theta_r} \tag{3.27}$$

- *The CDF of H_T*: This quantity is defined as $F_{H_T}(h) = Pr\left(S_R^+ \leq h \mid R > S_R^+\right)$. Due to the memoryless property of the exponential distribution $S_R^+ \equiv S$. Therefore $F_{H_T}(h)$ is expressed as the integration of the joint probability of S and R in the region limited by h and dividing the result by the probability that $(S \leq R)$ given as $Pr(S \leq R) = \int_0^\infty \int_0^y f_S(x)\,dx f_R(y)\,dy$. Following similar analysis as in (3.24), it can be shown that H_T is statistically equivalent to H_O as,

$$F_{H_T}(h) = Pr(S \leq h \mid S \leq R) = \frac{\int_0^\infty \int_0^{min(y,h)} f_S(x)\,dx f_R(y)\,dy}{Pr(S \leq R)} = F_{H_O}(h) \tag{3.28}$$

The proof of (3.28) is provided in Appendix A.1.3.

3.5.3.4 Mean Number of Accounting Interims and Reauthentications

To evaluate the mean number of re-authentications and interims during the session φ, we simply use (3.15) to calculate the respective means of interims and reauthentications within each AGW holding time period, φ_i, and multiply the corresponding results by the probability of occurrence of each holding time. Thus, from (3.15) we have,

$$\varphi_i(\Delta) = \sum_{n=1}^{\infty} \bar{F}_{H_i}(n\Delta) \quad , \quad i \in \{F, O, Tr, T\} \tag{3.29}$$

where we have $E[I_{H_i}] = \varphi_i(\Delta_T)$ for the interims and $E[\mathrm{Re}_{H_i}] = \varphi_i(\Delta_M)$ for reauthorizations. Since H_F occurs only if no handoffs occur (i.e., with probability p_0 defined in (3.35)) while other holding times occur with the complementary probability $(1 - p_0)$, the mean number of interim messages during the session are given as,

$$E[I] = p_0 E[I_{H_F}] + (1 - p_0) E\left[I_{H_O}\right] + (1 - p_0) E[I_{H_T}]$$
$$+ E(\text{No. transit AGWs} \mid \text{No. Handoffs } (K) > 1) \Pr\{K > 1\} E[I_{H_{Tr}}]$$
$$= p_0 E[I_{H_F}] + 2(1 - p_0) E\left[I_{H_O}\right] + E[\text{No. transit AGWs} \cap K > 1] E[I_{H_{Tr}}]$$

E (No. transit AGWs $\cap$ No. Handoffs $(K) > 1$) can be evaluated by noting that the number of transit AGWs is $K-1$ and hence,

$$E\,(\text{No. transit AGWs} \cap \text{No. Handoffs}\,(K) > 1) = E\,(K-1 \cap K > 1)$$

is given as,

$$\begin{aligned} E\,(K-1 \cap K > 1) &= \sum_{k=2}^{\infty} (k-1)\Pr\{K=k\} \\ &= E[K] - \Pr\{K=1\} - (1 - \Pr\{K=1\} - \Pr\{K=0\}) = E[K] - 1 + p_0 \end{aligned} \tag{3.30}$$

Thus, the mean number of interims, $E\,[I]$ is given as,

$$E\,[I] = p_0 E\,[I_{H_F}] + 2\,(1-p_0)\,E\left[I_{H_O}\right] + (E\,[K] - 1 + p_0)\,E\,[I_{H_{Tr}}] \tag{3.31}$$

Thus, $E[\xi_{Int}] = \sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)} E\,[I]$. The mean number of reauthentications is obtained in exactly the same manner as in (3.31) by replacing $E\,[I_{H_i}]$ by $E\,[\text{Re}_{H_i}]$. The final parameter left to evaluate (3.31), is the mean number of handoffs as discussed next.

3.5.3.5 The Number of Handoffs in a Session

In this section, we derive the density function, $Pr\,(K=k) = f_K\,(k)$, of the number of handoffs in a session S. It should be noted that more generalized results on this aspect were derived in several publications, most recently in [36, 106–108]. However, it is still instructive to derive a simplified result which can help our discussion. In Section 3.8, we make novel contributions towards modeling the mean number of handoffs handovers under more generalized assumptions of network size, roaming, mobility patterns, and users' distributions in the network.

Let $R_0^{(k)} = \tilde{R} + \sum_{j=1}^{k-1} R$ which denotes the sum of the residence times that a session incurs up to the k^{th} handoff moment. If we define $G\,(k) = Pr\left[S > R_0^{(k)}\right]$ (i.e., the probability that S includes at least k handoffs), then $G\,(k)$ is given as,

$$G\,(k) = \int_0^{\infty} \bar{F}_S\,(x) \left(f_{\tilde{R}}\,(x) \otimes \overbrace{f_R\,(x) \otimes .. \otimes f_R\,(x)}^{(k-1)^{th}-fold} \right) dx \tag{3.32}$$

Since the n^{th} fold convolution of Gamma density functions is Gamma distributed with shape and scale parameters of nk_r and θ_r respectively, then (3.32) can be expressed

using the Laplace transform of $R_0^{(k)}$ as,

$$G(k) = \frac{\theta_r \mathscr{L}\left\{\Gamma(k_r, y) \otimes \mathrm{PDF}_\gamma((k-1)k_r, y)\right\} \mid_{\hat{s}=\frac{\theta_r}{E_s}}}{\Gamma(k_r) E_r} = \frac{E_s}{E_r}\left(1 - \left(\frac{\theta_h}{\theta_r}\right)^{k_r}\right)\left(\frac{\theta_h}{\theta_r}\right)^{k_r(k-1)} \tag{3.33}$$

where θ_h is defined in (3.25). It follows that the probability that a session contains k handoffs (where $k \geq 1$) is written as,

$$\begin{aligned} f_K(k) &= Pr(K=k) = Pr\left(R_0^{(k+1)} > S \geq R_0^{(k)}\right) \\ &= Pr\left(S > R_0^{(k)}\right) - Pr\left(S > R_0^{(k+1)}\right) = G(k) - G(k+1) \end{aligned} \tag{3.34}$$

The probability of making no handoffs, p_0 is evaluated using (A.3) as,

$$\begin{aligned} p_0 &= Pr\left(S \leq \tilde{R}\right) = \int_0^\infty \int_0^y f_S(x)\, dx f_{\tilde{R}}(y)\, dy = \int_0^\infty F_S(x) f_{\tilde{R}}(x)\, dx \\ &= \int_0^\infty F_S(x) \frac{\bar{F}_R(x)}{E_r} dx = \int_0^\infty (1 - e^{-\frac{x}{E_s}}) \frac{\bar{F}_R(x)}{E_r} dx \\ &= 1 - \int_0^\infty e^{-\frac{x}{E_s}} \frac{\bar{F}_R(x)}{E_r} dx = 1 - \frac{\theta_r \mathscr{L}\{\Gamma(k_r, y)\} \mid_{\hat{s}=\frac{\theta_r}{E_s}}}{E_r \Gamma(k_r)} = 1 + \left(\left(\frac{\theta_h}{\theta_r}\right)^{k_r} - 1\right)\frac{E_s}{E_r} \end{aligned} \tag{3.35}$$

Thus by combining the results from (3.34) and (3.35), the PDF of the number of handoffs in a session is given as,

$$Pr(K=k) = \begin{cases} 1 + \left(\left(\frac{\theta_h}{\theta_r}\right)^{k_r} - 1\right)\frac{E_s}{E_r} & k = 0 \\ G(k) - G(k+1) & k \geq 1 \end{cases} \tag{3.36}$$

The mean number of handoffs can be written as,

$$E[K] = \sum_{n=0}^{\infty} nPr(K=n) = \sum_{n=1}^{\infty} G(K=n) = \frac{E_s}{E_r} \tag{3.37}$$

3.5.3.6 The Basic Model for AAA Signaling Load in Mobile Scenarios

The AAA signaling rate is given by substituting (3.21), (3.31) and (3.37) into (3.1) as,

$$E[\xi] = \sum_{i=1}^{N_{\mathrm{AGWs}}} \lambda^{(i)} \left[(1+2p_a)(E[K]+1) + p_a(E[I]+E[M])\right] \tag{3.38}$$

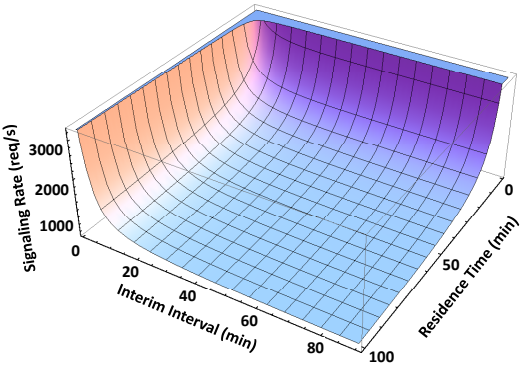


Figure 3.6: Mobility and interim interval effects on the mean AAA signaling rate in mobile networks (centralized AAA systems).

where $E[I]$ and $E[M]$ denote the mean number of interims and re-authentications in a session respectively. Fig.3.6 illustrates the effects of the AGW residence time and the accounting interim intervals on the mean signaling rate.

An approximation of (3.38) is obtained by assuming exponential residence times. Due to the memoryless property of the residence time R, $H_F \equiv H_O \equiv H_{Tr} \equiv H_T$ are all exponentially distributed. Hence, the number of interim messages in such periods can be found using (3.16). Therefore, the approximate signaling rate in (3.38) as,

$$E[\xi] = \sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)} (E[K]+1) \left[1 + p_a \left(2 + \frac{1}{e^{\frac{\Delta_T}{E_H}} - 1} + \frac{1}{e^{\frac{\Delta_M}{E_H}} - 1}\right)\right] \quad , \quad E_H = \frac{E_s}{E[K]+1} \tag{3.39}$$

3.5.4 Case Study: Mobility Profiles in an AGW Region

Although an AGW may cover a large region composed of a large number of cells (e.g., 100 cell sites), the users in the border region result in a considerable AAA signaling which is comparable to the majority of users who never leave the AGW region. To illustrate this aspect, let us consider two cellular topologies: 10×10 cells/AGW and 20×20 cells/AGW. For both topologies, we assume random movement patterns in four directions. The cellular residence times are lognormally distributed as they are known to fit real measurements [17, 109]. Since the cellular residence time is a result of both users' movements or cell reselection (e.g., due to signal fading effects or load balancing in the radio network), we also consider stationary users residing in the boundary zone between AGW regions.

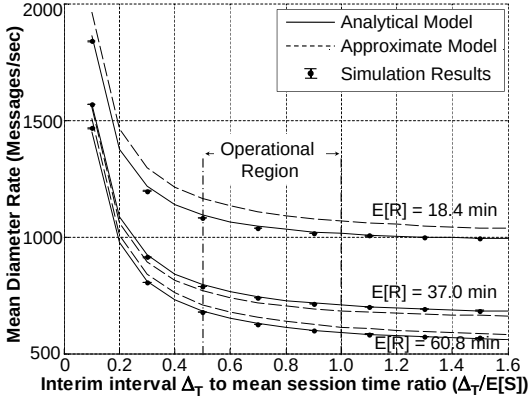


Figure 3.7: Interim interval effect on the mean AAA signaling rate in mobile networks [102]. Simulation parameters [5 AGWs, $\lambda^{(i)} = 20$ req/sec, E_S=Δ_M=40 min, 5×5 cells/AGW with residence times varying from 2.5-15 min/cell (Lognormal coeff. of var. $= 3$, model fit $C_R = 2$ for the whole area), mean batch method, 30 batches, 95% confidence (error bars within marker sizes)].

In our simulations, we assume three mean cellular residence times 5 and 10 mins. To be more conservative and to avoid exaggerating mobility effects, we only consider random mobility for users residing at most three cells away from the borders of the AGW, while the remaining users are assumed to be stationary (i.e., never leave the AGW region). In other words, the movement pattern is assumed to be random between cells and is limited within the outer three cellular rings. To gain a clear understanding of the residence time statistics, we categorize users into three types depending on their residence time (i.e., mobility profile): low, medium, and high. High mobility users constitute the lower 10th percentile of the residence time distribution and primarily represent users who are near the border. Medium mobility users are users whose speeds fall between the 10th and the 30th percentiles. Low mobility users correspond to the rest (i.e., the remaining 70% of the users). We also consider users who reside in the border cells in the overlapping regions with cells from neighboring AGWs. Since reselection can happen in the order of 2 mins [109], such users can lead to significant AAA signaling if they engaged in long file downloads or peer-to-peer file sharing activities extending over the whole day. We assume a fairly low density of such users. Table 3.2 summarizes our findings.

By observing the estimated AGW residence time statistics in Table 3.2, we see that the median (i.e., 50% of the residence time samples) is about fourth to sixth of the mean AGW residence time. This indicates heavy tail properties of the AGW residence time. This is clear from the coefficient of variation which is greater than one and ranges from 1.9-3.9 indicating that the standard deviation of the residence times is approximately double to quadruple its average. Notice that this AGW residence time is comparable to

Table 3.2: Access gateway residence times for different topologies and mobility profiles [100,000 samples, Lognormal cellular residence times coeff. of variation of 0.8].

		10 x 10 cells/AGW		20 x 20 cells/AGW	
Mobility Settings					
% Mobile Users to Total Users		84%		51%	
% Border Users to Total Users		0.1%		0.05%	
% Stationary Users to Total Users		15.9%		48.95%	
Estimated AGW Residence Time Statistics (in mins)					
Mean Cell Residence Time	**AGW Residence Time Statistics**				
5 mins	**Mean**	48.8		58	
	Median	14.6		9.6	
	Coeff. of Variation	1.9		3.7	
10 mins	**Mean**	98.3		113.4	
	Median	28.6		19.1	
	Coeff. of Variation	1.9		3.9	
Estimated Mobility Profiles Statistics (in mins)					
Mean Cell Residence Time	**Mobility Profile**	**Mean**	**Coeff. of Variation**	**Mean**	**Coeff. of Variation**
5 min	**Low**	68.3	1.5	81.8	3.1
	Medium	4.3	0.3	3.1	0.3
	High	1.3	0.5	1.0	0.5
10 min	**Low**	137.6	1.5	159.5	3.2
	Medium	8.5	0.3	6.3	0.3
	High	2.6	0.5	2.0	0.5
Estimated AAA Under-provisioning if Mobility is Ignored (i.e., 100% Stationary Users)					
5 min	**Percentage Under Provisioning**	168%		89%	
	Due to Mobile Users	103%		55%	
	Due to Border Users	65%		34%	
10 min	**Percentage Under Provisioning**	116%		61%	
	Due to Mobile Users	51%		27%	
	Due to Border Users	65%		34%	

the mean session duration that one would expect in broadband mobile networks (e.g., 30-60 mins). In addition, the number of cells per AGW region does not largely impact the mean residence time since increasing the AGW area also results in increasing the number of boundary cells (i.e., residence times in a 20x20 cells/AGW is around 1.2 times the residence time in a 10x10 cells/AGW region).

To attain a deeper understanding of the mobility statistics, we cluster our samples into three classes according to their mobility as described above (i.e., low (70% of the samples), medium (20% of the samples), and high (10% of the samples)). From Table 3.2,

we see that high mobility users incur an average AGW residence time of around 1.0-2.6 mins with a fairly reasonable coefficient of variation of 0.3-0.5. Medium mobility users also incur 3.0-8.0 min AGW residence time with a coefficient of variation of 0.3. Most of the variation is due to the 70% low mobility users who barely leave the AGW region (i.e., coefficient of variation 1.5-3.2).

Using our AGW residence time statistics, we show that ignoring mobility can lead to large under provisioning of the AAA system. In this regard, we compare results obtained by the basic AAA model presented in this section to the results from the AAA model for fixed networks which was derived in the previous section. From Table 3.2, we clearly see that the signaling load due to mobile users leads to a significant under provisioning of the AAA system if mobility is ignored. We also see that the tiny percentage of users residing in the border cells between AGWs cause large under provisioning which ranges from 34%-65%. This is because these users have long sessions (24 hours in our example) and fairly short residence times due to frequent cell reselections (i.e., 2 mins according to measurements from [109]). To sum up, we see that ignoring mobility in the AAA system design can easily lead to under provisioning of the AAA system.

3.5.4.1 Case Study: Effect of Session Dropping on AAA Signaling Rate

In some cases, sessions are dropped during AGW handoffs due to excessive handoff delays or due to other factors such as unavailability of wireless resources. Such effects can be generically incorporated in the model in (3.38), by using the likelihood function of session dropping, denoted as ρ. For instance, for excessively long handoff delays (i.e., longer than d_a time units), the session is dropped with a probability of $\rho = \Pr(d > d_a)$. The probability ρ can be obtained from available analytical models such as in [110] or from measurements, as the handoff delay highly depends on the access technology and the used handoff mechanisms. For instance, according to [111, 112], the handoff delay is given as the sum of the Alternative Point-to-Point (AltPPP) Sync, AltPPP Request, AltPPP Reply, and ICMPv6 Router Advertisement messages. In EVDO systems, observations showed air link latencies of 99 ms and a standard deviation of 48 ms [113]. Since each message is sufficiently small to fit in one radio frame, and assuming a typical 50 ms handoff delay at the EVDO layer, 10ms RTT delay between the AAA system and the AGW, and 10 ms for authentication, the resulting mean delay is 466ms with a standard deviation of 100ms[4]. Using moment matching and assuming a Gamma fit similar to (3.19), it is possible to have the fitting $\rho = \Gamma(k_0, d_a/\theta_0)/\Gamma(k_0), k_d = (466/100)^2 = 21.72, \theta_0 = 466/k_d = 21.46$.

By going back to Fig. 3.5 and considering each of the three cases separately, once for a complete session and another for an incomplete one using the session dropping probability ρ, the following can be derived according to the three cases on the figure.

[4]Standard deviation is obtained as the sum of the delay variance of the four messages (i.e., AltPPP Sync, AltPPP Request, AltPPP Reply, and ICMPv6 Router Adv.) as $\sqrt{4 \times 48^2} \approx 100$ ms.

- *Case 1*: no modification is needed since sessions are not dropped (no handoffs).
- *Case 2 (i.e., exactly one handoff)*: if the session is dropped, then only the first period H_O will occur with probability of: $\Pr\{K \geq 1\}\rho = G(1)\rho$. Otherwise we will have both periods (i.e., H_O, H_T or H_O, H_O since they are equivalent) occurring with a likelihood of: $\Pr\{K = 1\}(1-\rho) = (G(1) - G(2))(1-\rho)$.
- *Case 3 (multiple handoffs)*: for an incomplete session dropping at the m^{th} handoff, the period H_O is followed by $(m-1)$ H_{Tr} periods. This event happens with a probability of $\Pr\{K \geq m\}(1-\rho)^{m-1}\rho = G(m)(1-\rho)^{m-1}\rho$. For a session completing all the m handoffs, this probability is: $\Pr\{K = m\}(1-\rho)^m = (G(m) - G(m+1))(1-\rho)^m$.

Notice that the handoff PDF and CCDF are given in (3.33) and (3.36). Consequently, the PDF of the number of AGW handoffs in (3.37) is updated as,

$$\Pr\{K = k\} = f_K(k) = (G(k) - G(k+1))(1-\rho)^k + G(k)(1-\rho)^{k-1}\rho \tag{3.40}$$

Therefore the effective number of handoffs $E[K_e]$ with session dropping is,

$$E[K_e] = \sum_{k=1}^{\infty} k f_K(k) = \frac{E_s}{E_r} \frac{1 - \left(\frac{\theta_h}{\theta_r}\right)^{k_r}}{1 - \left(\frac{\theta_h}{\theta_r}\right)^{k_r}(1-\rho)} \tag{3.41}$$

The last period H_T occurs only if the session is not dropped with a probability, p_l, as,

$$p_l = \sum_{k=1}^{\infty} (G(k) - G(k+1))(1-\rho)^k = \left(1 - \left(\frac{\theta_h}{\theta_r}\right)^{k_r}\right) E[K_e] \tag{3.42}$$

Since for k handoffs, we have $(k-1)$ H_{Tr} periods, the mean number of H_{Tr} periods is,

$$E[N_{H_{Tr}}] = \sum_{k=2}^{\infty} (k-1) f_K(k) = E[K_e] - (1 - p_0) \tag{3.43}$$

Thus, similar to (3.31) the mean number of interim messages in a session that can be potentially dropped is given as,

$$E[I] = p_0 E[I_{H_F}] + (1 - p_0 + p_l) E[I_{H_O}] + E[N_{H_{Tr}}] E[I_{H_{Tr}}] \tag{3.44}$$

The mean number of reauthentications, $E[M]$, can be evaluated similar to $E[I]$ by using Δ_M instead of Δ_T when evaluating $E[I_{H_F}]$, $E[I_{H_O}]$, and $E[I_{H_{Tr}}]$ respectively. Thus substituting (3.44) and (3.41) into (3.1), the mean AAA signaling rate which considers the

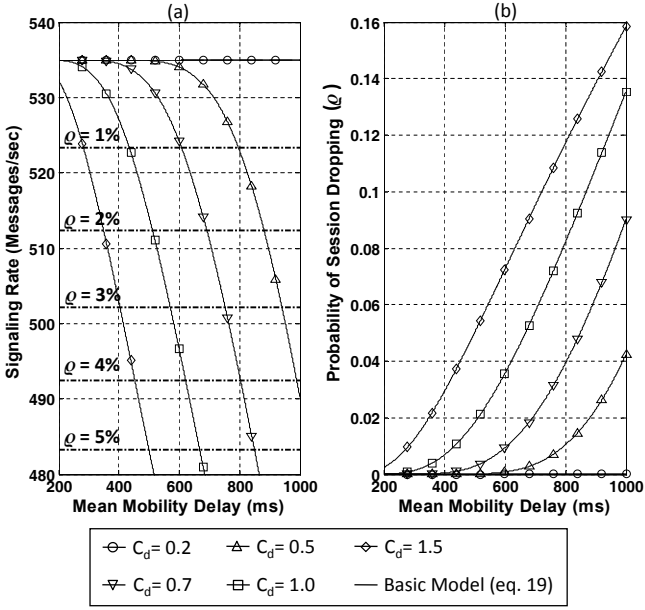


Figure 3.8: Signaling rate vs mean handoff delay [Parameters: E_s = 40 min, Δ_T=20 min, Δ_M = 40 min, E_R = 18.4 min, C_R = 2, p_a=0.97, d_a = 2.0 sec, c_d = 1.3, λ = 10 req/s, N_{AGW} = 5]

potential session dropping is given as,

$$E[\xi] = \sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)} \left[(1+2p_a)(E[K_e]+1) + p_a(E[I]+E[M]) \right] \tag{3.45}$$

We study the effect of the mean handoff delay as well as the variance of the delay characterized by the coefficient of variation C_d using Gamma fits. The sessions are dropped if the handoff signaling delay exceeds the maximum delay, d_a, as d_a>2s. Figures 3.8.(a)-3.8.(b) show the effect of the handoff signaling delay on the resulting AAA signaling rate and the the session dropping probability. We observe that as the handoff delay and the session dropping probability increase, the corresponding AAA signaling rate decreases. We also see that highly varying handoff delays (i.e., large C_d) result in higher session dropping. For nominal session dropping rates of ($< 2\%$), the resulting AAA signaling rate can be approximated by the model in (3.38) instead of (3.45).

3.5.5 Basic AAA Mobile Network Model's Limitations

In addition to some of the limitations for the fixed rate model relevant to the session arrival process, the processing power, and the users' quota, the following are additional limitations that can cause errors in the predicted AAA rate.

1. *The homogeneous residence time assumption*: We use this simplifying assumption for tractability. This assumption is followed almost in all analytical research work in the area of mobility in cellular networks such as in [17, 36, 37, 103]. In the next section, we relax this assumption.

2. *The exponential session duration assumption*: This is needed to simplify the derivation of the AGW holding times. The mean number of handoffs $E[K]$ in our model applies for general distributions as argued in [36, 106–108, 114]. For session coefficient of variations less than or equal to 2, we showed that the model's estimation error is within practical limits and is below 15% (see Section 5.2).

3. *The knowledge of the AGW residence time*: The model assumes that the AGW residence times are known. This assumption might be limiting when all available measurements are based on cellular residence times. In Section 3.8.4, we show generic means of deriving the AGW residence time using the cellular residence time measurements.

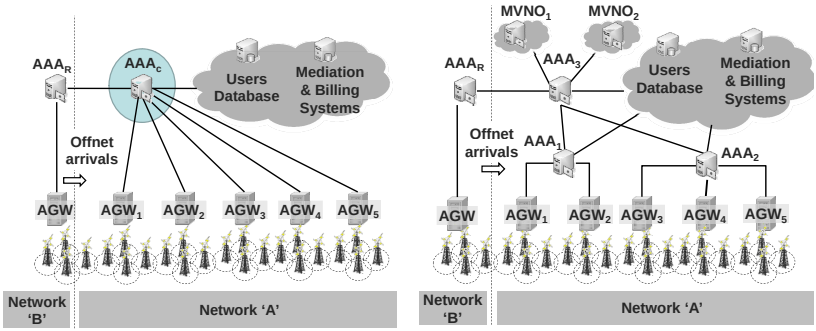
4. *Other AAA configurations*: The model only handles centralized AAA deployments where the AAA system receives all the signaling pertaining to a session irrespective of the user's serving network or AGW. As such it does not handle the cases of multiple AAA systems in the network where the AAA might be handling only parts of the session due to mobility within the network or roaming to other operators' networks. We address this important case in the next section.

3.6 Distributed AAA Systems and Roaming Users

The model presented so far in (3.38) only captures the signaling rate at the AAA system in centralized deployments. There are many reasons why operators would use a decentralized AAA system especially to expand their networks to support broadband data applications. This includes minimization of signaling delay by deploying AAA systems in closer geographic proximity to the AGW locations, load balancing in the network, geographic redundancy, cost, etc. The models we have studied so far do not address the signaling pertaining to roaming users where even a centralized AAA system in the network may not receive *all* signaling traffic pertaining to the session as the user may roam into other networks. Roaming traffic is of interest as it is expected to increase significantly due to the growing adoption of Mobile Virtual Network Operators (MVNOs)

models[5], where third parties can offer mobile services without owning wireless network infrastructure [24, 25, 115].

To illustrate the planning problem, consider the exemplary centralized AAA system in Fig. 3.9(a). The network includes five AGWs served by the AAA system as well as other supporting systems such as mediation and billing systems and user databases. The AAA system serves home and roaming AAA signaling requests from the five AGW regions. Roaming users can belong to MVNO partners or to Network 'B'. Suppose that the starting locations of sessions are equally likely in all AGWs (i.e., 20% each). Now let us consider the case where the operator decides to split the load on the central AAA into two distributed AAA systems in the network as shown in Fig.3.9(b); AAA1 serves AGWs 1-2 and AAA2 serves the rest. AAA1 and AAA2 directly handle AAA signaling requests for home users while they forward all requests pertaining to roaming users to AAA3 for processing and routing to their provider networks. On the first thought, one may think that AAA3 will behave similarly to the centralized system for roaming users while AAA1 will handle 40% of the signaling load for home users while AAA2 will handle the rest. However, deeper consideration of mobility reveals that the mobility pattern among AGWs as well as the session statistics can shift the load from one AGW into another. Clearly, mobility leads to non-uniform splitting of the signaling load pertaining to a session between AGWs and hence their serving AAAs in the network. This leads us to the interesting question that given the range of AAA deployment choices and effects of mobility, would it still be possible to plan AAA systems analytically without resorting to exhaustive simulations ?



(a) Centralized System (b) Distributed AAA System Deployment

Figure 3.9: Exemplary centralized and distributed AAA system deployments.

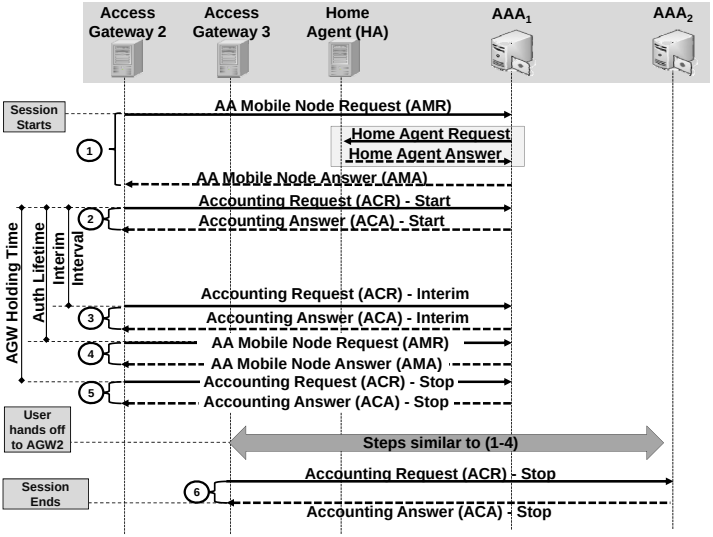
[5] According to [26, 27], the MVNO market is expected to grow from $4 billion in 2005 to top $25 billion by 2012 covering hundreds of millions of users; also the telemetry machine-to-machine market facilitated by MVNO models is expected to grow from $15 billion in 2008 to $57 billion in 2014.

In this section, we devise a generic planning methodology that can handle both centralized and distributed AAA systems even in the presence of roaming and mobility. Up to our knowledge, this is the first effort in the literature that addresses this issue for designing AAA systems. In our approach, we use a generic framework that allows the calculation of the signaling load on centralized and distributed AAA systems. The generalized framework is able to handle cases where AGWs can have different residence times (i.e., mobility profiles). It can also handle cases with different authentication protocol and accounting settings per AGW which was not studied so far neither for centralized nor distributed systems. The proposed approach utilizes stochastic and renewal theoretic concepts combined with transient Markov chains to address protocol and user movement behaviors. We consider relevant aspects including AAA protocol settings, mobility, session duration, user distribution on AGW areas, and AGW residence times.

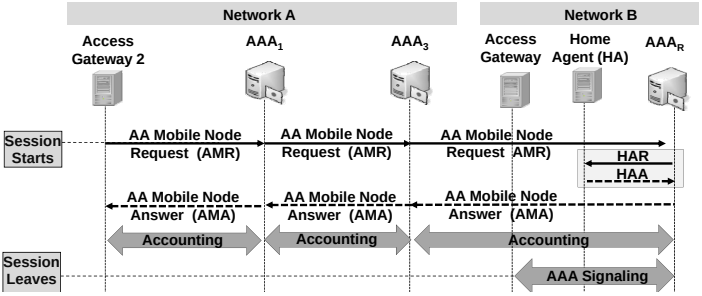
3.6.1 Reference Architecture

Based on Fig.3.9(b), Fig.3.10(a) shows a typical AAA signaling diagram for home users (i.e., Network A) and using the terminology from the Diameter protocol [45, 61]. The process is identical to the one shown in Fig. 3.4 and is repeated to illustrate how the signaling occurs when more than one AAA system serves the session. When a mobile node initially accesses the network, it sends a Mobile IP registration request to the serving AGW (i.e., AGW_2 in Fig.3.10). This triggers an authentication request (AA-Mobile-Node-Request (AMR)) from AGW_2 to AAA_1 (step 1). AAA_1 authenticates the user. Based on [61], it is possible to have an extra signaling message, called Home Agent Request (HAR) which facilitates Mobile IP registration and establishes security association between the mobile node and the home agent is sent from AAA_1 to the home agent. Upon receiving the home agent answer (HAA), the AAA system returns an AA-Mobile-Node-Answer (AMA) to the requesting AGW. Afterwards, the session is started and AGW_2 sends an accounting request ACR(Start) towards AAA_1 indicating the initiation of the session (step 2). Afterwards, AGW_2 periodically sends ACR(Interim) messages every Acct-Interim-Interval to report the time and volume usage during the service lifetime (step 3). When the Authorization-Lifetime is about to expire, the mobile node sends a Mobile IP registration request to AGW2 which triggers an AMR message to the AAA server (step 4). Hence, the session is periodically re-authenticated once the Authorization-Lifetime elapses. If the subscriber moves (i.e., hands off) between AGWs (i.e., AGW_2 to AGW_3 here), an ACR(Stop) is sent by AGW_2 to AAA_1 (step 5) while an AMR followed by an ACR(Start) messages are sent by the AGW_3 to AAA_2. Afterwards, periodic interims and reauthentications are sent by AGW_3. An ACR(Stop) is sent when the session terminates.

For roaming users, the signaling process is quite similar as shown in Fig.3.10(b) with the difference that now AAA_1 forwards authentication and accounting requests towards 'Network B'. When enabled, the home agent signaling (i.e., the HAR/HAA exchange) is usually carried in the home network (i.e., Network B) as in Fig.3.10(b). If the roam-



(a) AAA signaling for home users



(b) AAA signaling for roaming users

Figure 3.10: AAA signaling for home and roaming users

ing user moves back to their home network, the serving AGW in 'Network B' interacts directly with AAA_R without involving AAA systems from 'Network A'. It should be noted that the home network may choose to let the visited network (i.e., 'Network A') dynamically assign the home agent. In this case, AAA_R in 'Network B' forwards the HAR requests towards AAA3 which can carry the home agent assignment from 'Network A'.

3.6.2 The Generalized AAA Planning Model

Our goal in this subsection is to demonstrate how to design centralized and distributed AAA systems in mobile networks in a conceptually similar manner as load splitting (or balancing) in fixed networks. This is achieved by designing the system similar to the basic model in Section 3.5 while at the same time keeping track of the load coming from each AGW. This is achieved by combining a transient Markov chain, which tracks mobility at each AGW, with our stochastic model in Section 3.5. A brief summary of relevant results of transient Markov chains is available in Appendix B.1. For the sake of discussion, we briefly repeat some key topics from Section 3.5 especially those relevant to the AGW holding time. Furthermore, in our analysis, we call the session arrivals that start from within the network as on-net sessions whereas sessions that start in other networks and then handoff to the network under consideration at a random point during their lifetime as off-net sessions. Again, we use the term handoffs to refer to the movement between AGW regions and not between cells.

3.6.2.1 Assumptions

In addition to the assumptions relevant to the session and residence times in Section 3.5.2, we have the following assumptions in our analysis,

1. The *on-net* and *off-net* session arrival processes are Poissonian with mean rates of λ_Ω, and λ_Φ respectively.
2. The mobility model among AGWs is assumed to be Markovian similar to [114, 116–119]. This means that once the residence time elapses, the users may move in all possible directions with likelihoods according to the mobility model.

3.6.2.2 AAA Signaling

In this section, we develop a generic formulation for AAA signaling which we use later to obtain the AAA signaling load for centralized and distributed AAA deployments. Let us denote the number of authentication and reauthentication messages during a session $E[\xi_A]$ and $E[\xi_{Re}]$ respectively. Let us also denote the number of accounting start, interim, and stop messages as $E[\xi_{Start}]$, $E[\xi_I]$, and $E[\xi_{Stop}]$ respectively. Let λ denote the total session arrival rate in the network. Since accounting messages are sent for already authenticated messages and assuming that reauthentications are always successful after the initial authentication, it follows that the AAA signaling load can be written as the sum of all AAA messages from all access gateways as,

$$E[\xi] = \lambda \left[E[\xi_A] + p_a \left(E[\xi_{Start}] + E[\xi_{Stop}] + E[\xi_I] + E[\xi_{Re}] \right) \right] \tag{3.46}$$

where p_a denotes the authentication success probability. From Fig.3.10(a), for each AGW, only one authentication operation is performed followed by an accounting start and later by an accounting stop message. Since for one handoff we have two authentications and accounting start and stop messages from the two AGWs, then the aggregate number of such messages during the session's lifetime from all AGWs for K handoffs is proportional to the mean number of incurred handoffs plus one if the session initiates in the network, (i.e., on-net sessions (Ω)), $E[K]$, as,

$$E[\xi_A] = \delta(E[K] + \mathbb{I}[\Omega]) \quad , \quad E[\xi_{Start}] = E[\xi_{Stop}] = p_a(E[K] + \mathbb{I}[\Omega]) \tag{3.47}$$

where $\mathbb{I}[\Omega]$ is an indicator function and is equal to 1 if the session is initiated in the network otherwise it equals 0. Notice that $\delta = 1$ when only the AMR exchange is handled by the AAA system (see AAA1 in Fig. 3.10(b)), and $\delta = 2$ when (AMR and HAR) exchanges are handled by the AAA system (see AAA1 in Fig. 3.10(a)).

Since depending on the protocol settings of the interim interval, Δ_T, and authorization lifetime, Δ_M, more than one interim and reauthentication messages can be sent by the serving AGW. This is a function of the time, H, a session spends within an AGW region. We refer to the time, H, as the AGW holding time. Conceptually there are four types of AGW holding times depending on the initiation and termination instants of the session from the perspective of the serving AGW as follows.

- *Full Session, H_F* : In this case, only one AGW serves the whole session. The CDF of the AGW holding time, H_F, is given in (3.24).
- *Originating Session, H_O* In this case, the AGW serves a session that starts within its coverage then leaves to other AGW regions. The CDF of H_O is given in (3.26).
- *Transit Sessions, H_{Tr}*: In this case, the AGW serves a transiting session which already started somewhere else and terminates in another AGW area. The CDF of H_R is given in (3.27).
- *Terminating Sessions, H_T*: In this case, the AGW serves a session that starts somewhere else and then terminates within its coverage area. The distributions of H_O and H_T are statistically equivalent under the exponential session assumption.

Thus, the mean number of interim ($E[I_{H_i}]$) and reauthentication ($E[Re_{H_i}]$) messages per AGW can be calculated by taking the expectation of the floor of the random variable H_i and Δ as $E\left[\lfloor \frac{H_i}{\Delta} \rfloor\right]$ and substituting the interim interval Δ_T or the authorization lifetime Δ_M for Δ accordingly as,

$$E[I_{H_i}] = E\left[\lfloor \frac{H_i}{\Delta_T} \rfloor\right] \quad , \quad E[Re_{H_i}] = E\left[\lfloor \frac{H_i}{\Delta_M} \rfloor\right] \quad , \quad E\left[\lfloor \frac{H_i}{\Delta} \rfloor\right] = \sum_{j=1}^{\infty} \bar{F}_{H_i}(j\Delta) \tag{3.48}$$

Hence, the aggregate number of the interims/reauthentications received by all AGWs in the network depends on the sequence of the holding times incurred by the mobile device as it moves in the network. For instance, a mobile residing in the border region between AGWs 1-3, maybe served by AGW1 then AGW2 then terminate in AGW3 resulting in the ordered sequence of AGW holding times as $\{H_{O,1}, H_{R,2}, H_{T,3}\}$. Another example, a roaming user may initiate their session in the visited network and terminate it in other networks resulting in AGW holding time sequence as, $\{H_{O,1}, H_{R,2}\}$. Hence, the expected number of interims/reauthentications is obtained by taking the average of their quantities, $N_{H_{i,j}}$, over all possible holding time categories $i \in \{F, O, Tr, T\}$, from all AGWs as,

$$E[\xi_I] = \underset{\forall H_{i,j}}{\mathrm{E}} \left[\sum_{n=1}^{N_{H_{i,j}}} I_{H_{i,j}} \right] \quad , \quad E[\xi_{Re}] = \underset{\forall H_{i,j}}{\mathrm{E}} \left[\sum_{n=1}^{N_{H_{i,j}}} Re_{H_{i,j}} \right] \tag{3.49}$$

Now that we have defined all the quantities in (3.46), we proceed to show how we can calculate the mean number of handoffs, K, in (3.47) as well as the expectations in (3.49). To do so, we first introduce few concepts relevant to mobility modeling and then derive each of these quantities accordingly.

3.6.2.3 Mobility

In our analysis, we assume a Markovian mobility model where depending on the serving AGW (AGW_j), the user hands off to a target AGW (AGW_k) according to a movement probability of m_{jk}. The handoff occurs if the remaining session duration is longer than the residence time, R_j, of the mobile within AGW_j, otherwise the session terminates. Notice that the remaining session duration is statistically equivalent to the session duration due to the memoryless property of the session distribution. Thus, a handoff from AGW j to k occurs if the mobile chooses to move to AGW k and if the session duration is greater than the residence time in AGW j (i.e, with a probability of $m_{jk}\Pr\{S > R_j\}$). Let us define the elements of the matrix $\mathbf{Q}$ as $\|\mathbf{Q}\|_{jk} = m_{jk}\Pr\{S > R_j\}$ where the $\Pr\{S > R_j\}$ can be obtained using the Laplace transform of the residence time as,

$$l_j = \Pr\{S > R_j\} = \mathscr{L}\{R_j\} \mid_{\hat{s}=E_s^{-1}} = \left(\frac{1}{1 + \hat{s}\theta_r^j} \right)^{k_r^j} \mid_{\hat{s}=E_s^{-1}} = \left(\frac{E_s}{E_s + \theta_r^j} \right)^{k_r^j} \tag{3.50}$$

where k_r^j and θ_r^j are the shape and scale parameters of the gamma fit of R_j. We refer to $\mathbf{Q}$ as the *handoff matrix*. Since new sessions start from within the AGW area, they do not incur the full residence time R_j but rather its residual lifetime, denoted as $\tilde{R}_j$ [36, 102]. Thus, for on-net sessions, if $S > \tilde{R}_j$ the first handoff occurs, otherwise for

subsequent handoffs, $S > R_j$. Defining $l'_j = \Pr\{S > \tilde{R}_j\}$ we get,

$$l'_j = \Pr\{S > \tilde{R}_j\} = \mathscr{L}\{\tilde{R}_j\}\,|_{\hat{s}=E_s^{-1}} = \frac{1-\mathscr{L}\{R_j\}}{\hat{s}E_r}\,|_{\hat{s}=E_s^{-1}} = \frac{E_s}{E_r}(1-l_j) \tag{3.51}$$

For off-net sessions, since by definition, they made at least one handoff to enter the network under consideration, handoffs occur if $S > R_j$. To facilitate the discussion, we denote the handoff matrix for onnet sessions as $\mathbf{Q^{(\Omega)}}$ and for offnet sessions as $\mathbf{Q^{(\Phi)}}$. Thus, we formulate the handoff matrices for on-net and off-net traffic as,

I. For on-net traffic: The handoff matrix, $\mathbf{Q}^{(\mathbf{\Omega})}$, is expressed by the $2N_{AGW} \times 2N_{AGW}$ matrix as,

$$\mathbf{Q}^{(\mathbf{\Omega})} = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{0} & \mathbf{B} \end{bmatrix} \tag{3.52}$$

where $\mathbf{0}$ is a $N_{AGW} \times N_{AGW}$ zeros matrix, and $\mathbf{A}$ and $\mathbf{B}$ are $N_{AGW} \times N_{AGW}$ matrices as follows,

$$\mathbf{A} = \begin{bmatrix} 0 & l'_1 m_{12} & l'_1 m_{13} & \cdots & & \\ l'_2 m_{21} & 0 & l'_2 m_{23} & l'_2 m_{24} & \cdots & \\ l'_3 m_{31} & l'_3 m_{32} & 0 & l'_3 m_{34} & l'_3 m_{35} & \cdots \\ & & \vdots & \ddots & \vdots & \\ l'_{N_{AGW}} m_{N_{AGW}1} & \cdots & & & l'_{N_{AGW}} m_{N_{AGW}(N_{AGW}-1)} & 0 \end{bmatrix} \tag{3.53}$$

The matrix $\mathbf{B}$ is defined similarly to $\mathbf{A}$ by replacing l'_j by l_j.

II. For off-net traffic: The transition probabilities are similar to the on-net case with the difference that all handoffs occur if $S > R_j$. Hence, using (3.52)-(3.53), it follows that the handoff matrix, $\mathbf{Q}^{(\mathbf{\Phi})}$, is defined only using an $N_{AGW} \times N_{AGW}$ matrix as,

$$\mathbf{Q}^{(\mathbf{\Phi})} = \mathbf{B} \tag{3.54}$$

The last interesting case is the description of the mobility for roaming users. Since roaming users only trigger signaling while in the visited network (i.e., the network under consideration), we need to consider the network departure to other operators, denoted as Z, from border AGWs. To do so, we define the roaming column vectors for on-net and off-net sessions as $\mathbf{U_Z}^{(\mathbf{\Omega})}$ and $\mathbf{U_Z}^{(\mathbf{\Phi})}$ respectively as,

$$\mathbf{U_Z}^{(\mathbf{\Omega})} = \begin{bmatrix} \mathbf{A_Z} \\ \mathbf{B_Z} \end{bmatrix} \quad , \quad \mathbf{U_Z}^{(\mathbf{\Phi})} = \mathbf{B_Z} \tag{3.55}$$

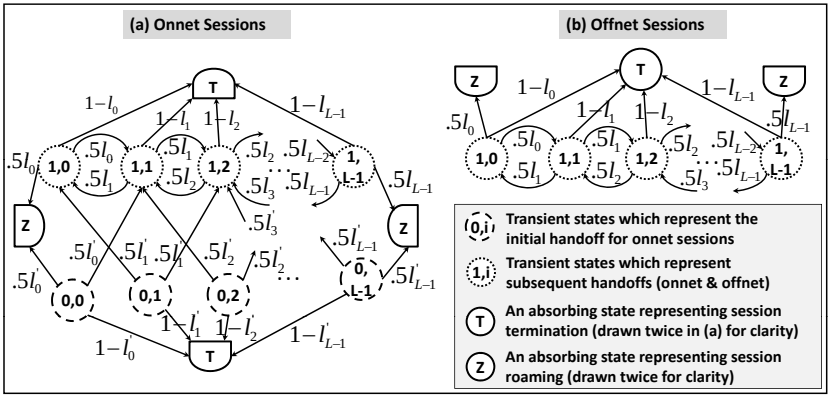


Figure 3.11: Exemplary transient Markov chain model for random mobility.

where $\mathbf{A_Z}$ and $\mathbf{B_Z}$ are $N_{AGW} \times 1$ column vectors as,

$$\mathbf{A_Z} = \begin{bmatrix} m_{1Z} l'_1 \\ m_{2Z} l'_2 \\ \vdots \\ m_{N_{AGW}Z} l'_3 \end{bmatrix}, \mathbf{B_Z} = \begin{bmatrix} m_{1Z} l_1 \\ m_{2Z} l_2 \\ \vdots \\ m_{N_{AGW}Z} l_3 \end{bmatrix} \tag{3.56}$$

Thus, in roaming situations users can move not only within AGWs in the network but also to other networks and hence their mobility is described using the handoff matrices $\mathbf{Q}$ and the roaming vector $\mathbf{U_Z}$. Therefore, in general, at a handoff instant, users may handoff to an AGW within the network or roam to another operator. Otherwise, their session would have terminated in the current AGW region. Thus, the vectors of the termination probabilities at each AGW in the network for on-net and off-net traffic are,

$$\mathbf{T^{(\Omega)}} = \begin{bmatrix} \mathbf{A_T} \\ \mathbf{B_T} \end{bmatrix} \quad , \quad \mathbf{T^{(\Phi)}} = \mathbf{B_T} \tag{3.57}$$

where $\|\mathbf{A_T}\|_j = 1 - \sum_k \|\mathbf{A}\|_{(j,k)} - \|\mathbf{A_Z}\|_j = 1 - l'_j$ and $\|\mathbf{B_T}\|_j = 1 - \sum_k \|\mathbf{B}\|_{(j,k)} - \|\mathbf{B_Z}\|_j = 1 - l_j$.

In the following two subsections, we use transient Markov chains theory in [120, 121] to obtain the mean number of handoffs as well as the termination and roaming probabilities. In our analysis, we view session termination and roaming to other networks as absorbing states while being served by an AGW in the network under consideration as transient states. Fig.3.11 illustrates an exemplary L linearly arranged AGWs with equal likelihoods of movement between adjacent AGWs in the east and west directions. Other arrangements can also be accommodated in a straight forward manner.

3.6.2.4 The Mean Number of AGW Handoffs

Hence, the mean number of handoffs in the network can be characterized as the mean number of state visits excluding the initial visit which corresponds to the session initiation event in the first AGW's area. To obtain the number of state visits, we use the fundamental matrix definition [120, 121], $\mathbf{M_q}$, as,

$$\mathbf{M_q^{(x)}} = \left(\mathbf{e} - \mathbf{Q^{(x)}}\right)^{-1} \quad , \quad \mathbf{x} \in \{\Omega, \Phi\} \tag{3.58}$$

where $\mathbf{e}$ is an identity matrix of the size[6] of $\mathbf{Q^{(x)}}$. Let distribution of the session initiation in the AGWs be given by the $1 \times N_{AGW}$ row vector: $\mathbf{F_I^{(\Omega)}}$ for on-net, and $\mathbf{F_I^{(\Phi)}}$ for off-net sessions. The elements of $\mathbf{F_I^{(\Phi)}}$ usually reflect the load from border AGWs where off-net traffic emerges (i.e., zeros are used for interior AGWs). On the other hand, $\mathbf{F_I}^{(\Omega)}$ reflects session initiation probabilities from all AGWs. Thus, the mean *number of visits* ($\|\boldsymbol{\nu}_{\mathbf{G}}^{(\mathbf{x})}\|_j$) to each AGW before roaming or session termination for on-net and off-net sessions is given by the row vector,

$$\boldsymbol{\nu}_G^{(\mathbf{x})} = \mathbf{P_I^{(x)}} \mathbf{M_q^{(x)}} \mathbf{D^{(x)}} \quad , \quad \mathbf{x} \in \{\Omega, \Phi\} \tag{3.59}$$

where $\mathbf{P_I^{(x)}}$ is the initial probabilities row vector and is defined as $\mathbf{P_I^{(\Omega)}} = [\mathbf{F_I^{(\Omega)}} \mathbf{0}]$ and $\mathbf{P_I^{(\Phi)}} = \mathbf{F_I^{(\Phi)}}$. The matrix $\mathbf{D^{(\Omega)}} = \begin{bmatrix} \mathbf{e} \\ \mathbf{e} \end{bmatrix}$ where $\mathbf{e}$ is an $N_{AGW} \times N_{AGW}$ identity matrix and is used to sum the two cases: initial handoffs which occur if $(S > \tilde{R})$ and the subsequent ones which occur if $(S > R)$ for each AGW. $\mathbf{D^{(\Phi)}} = \mathbf{e}$ and is only used for notation consistency. Let us denote the sum of the N_{AGW} elements of $\boldsymbol{\nu}_G^{(\mathbf{x})}$ row vector as $\nu_\Omega = \sum_i^{N_{AGW}} \|\boldsymbol{\nu}_G^{(x)}\|_i$. Then, the mean number of handoffs at each AGW for on-net traffic (K_Ω) is given by ν_Ω minus 1 to exclude the initial visit for session initiation. On the other hand, (K_Φ) for off-net sessions equals ν_Φ as the session has already been initialized outside the network. In vector form for the mean number of handoffs at each AGW we have,

$$\mathbf{K_G^{(\Omega)}} = \mathbf{P_I}^{(\Omega)} (\mathbf{M_q}^{(\Omega)} - \mathbf{e}) \mathbf{D}^{(\Omega)} \quad , \quad \mathbf{K_G^{(\Phi)}} = \mathbf{P_I}^{(\Phi)} \mathbf{M_q}^{(\Phi)} \tag{3.60}$$

Therefore, the total number of handoffs in the network from all AGWs for onnet and offnet sessions, denoted as $E[K_x]$, is obtained by summing the individual components of $\mathbf{K_G^{(x)}}$ in (3.60) by multiplying it by an $N_{AGW} \times 1$ all ones column vector ($\mathbf{o}$), as,

$$E[K_x] = \mathbf{K_G^{(x)}} \mathbf{o} \tag{3.61}$$

Thus, the number of authentications as well as accounting start and stop messages for on-net and off-net sessions in (3.47) can now be obtained from all AGWs using the

[6] We use $\mathbf{e}$ instead of the common symbol $\mathbf{I}$ to avoid confusion with the number of interims, I.

$E[K_x]$ or from each AGW (e.g., AGW$_j$) using $\boldsymbol{\nu}_G^{(\mathbf{x})}$. Finally, it is noteworthy to mention that the matrix $\mathbf{M_q}^{(\mathbf{\Omega})}$ can be simplified as,

$$\mathbf{M_q}^{(\mathbf{\Omega})} = \begin{bmatrix} \mathbf{e} & \mathbf{A}(\mathbf{e}-\mathbf{B})^{-1} \\ \mathbf{0} & (\mathbf{e}-\mathbf{B})^{-1} \end{bmatrix} \tag{3.62}$$

This can be verified by checking that $\mathbf{M_q}^{(\mathbf{\Omega})}\left(\mathbf{e}-\mathbf{Q}^{(\mathbf{\Omega})}\right)=\mathbf{e}$. By definition of $\mathbf{Q}^{(\mathbf{\Phi})}$, the matrix $\mathbf{M_q}^{(\mathbf{\Phi})}$ is given as $\mathbf{M_q}^{(\mathbf{\Phi})}=(\mathbf{e}-\mathbf{B})^{-1}$.

3.6.2.5 The Roaming and Session Termination Likelihoods

We now derive the likelihoods of leaving the network under consideration (i.e., roaming probability) and session termination. We use these terms to derive the number of interim and reauthentication messages later in this section. In our analysis, we derive the roaming probability β_x and the termination probability α_x for the whole network for on-net and off-net sessions. We also derive the likelihoods of roaming from each AGW, represented by the elements of the row vector $\|\boldsymbol{\beta}_G^{(x)}\|_i$, and session termination at each AGW, represented by the elements of the row vector $\|\boldsymbol{\alpha}_G^{(x)}\|_i$. Since the sum of the elements of $\boldsymbol{\alpha}_G^{(x)}$ and $\boldsymbol{\beta}_G^{(x)}$ leads to α_x and β_x, we refer to the latter as the total probabilities. We also refer to $\boldsymbol{\alpha}_G^{(x)}$ and $\boldsymbol{\beta}_G^{(x)}$ as the vector forms.

Using the transient Markov chain theory, it can be shown that the roaming probabilities for each AGW, denoted as $\boldsymbol{\beta}_G^{(x)}$, can be written as,

$$\boldsymbol{\beta}_G^{(x)} = \mathbf{P_I}^{(\mathbf{x})}\mathbf{M_q}^{(\mathbf{x})}\text{diag}\left\{\mathbf{U_Z}^{(\mathbf{x})}\right\} \quad , \quad x \in \{\Omega, \Phi\} \tag{3.63}$$

where diag$\left\{\mathbf{U_Z}^{(\mathbf{x})}\right\}$ operation places the elements of the vector $\mathbf{U_Z}^{(\mathbf{x})}$ on the diagonal of an identity matrix of the same length as $\mathbf{U_Z}^{(\mathbf{x})}$. The roaming probability from all AGWs is simply the sum of the components of $\boldsymbol{\beta}_G^{(x)}$ and is given as,

$$\beta_x = \boldsymbol{\beta}_G^{(x)}\mathbf{o} = \mathbf{P_I}^{(\mathbf{x})}\mathbf{M_q}^{(\mathbf{x})}\mathbf{U_Z}^{(\mathbf{x})} \tag{3.64}$$

Notice that the total roaming probability in (3.64) is calculated similarly to (3.63) but without using the diag$\{.\}$ operation. This fact applies to all results in this section. Substituting (3.62) into (3.64), it is easy to show that,

$$\boldsymbol{\beta}_G^{\Omega} = \mathbf{F}^{\Omega}\text{diag}\{\mathbf{A_z}\} + \mathbf{F}^{\Omega}\mathbf{A}(\mathbf{e}-\mathbf{B})^{-1}\text{diag}\{\mathbf{B_z}\}, \quad \boldsymbol{\beta}_G^{\Phi} = \mathbf{F}^{\Phi}(\mathbf{e}-\mathbf{B})^{-1}\text{diag}\{\mathbf{B_z}\} \tag{3.65}$$

where the term $\mathbf{F}^{\Omega}\text{diag}\{\mathbf{A_z}\}$ indicates leaving the network under consideration at the first handoff instant from each AGW while the $\mathbf{F}^{\Omega}\mathbf{A}(\mathbf{e}-\mathbf{B})^{-1}\text{diag}\{\mathbf{B_z}\}$ term indicates

making at least one handoff inside the network then leaving from each AGW. Again, expressions for β_Ω and β_Φ can be obtained by removing the diag $\{.\}$ operation in (3.65).

Finally, the session termination probabilities from each AGW, $\boldsymbol{\alpha}_{\boldsymbol{G}}^{(x)}$, are derived similarly to the roaming probabilities, $\boldsymbol{\beta}_{\boldsymbol{G}}^{(x)}$ as,

$$\boldsymbol{\alpha}_{\boldsymbol{G}}^{(x)} = \mathbf{P}_{\mathbf{I}}^{(\mathbf{x})}\mathbf{M}_{\mathbf{q}}^{(\mathbf{x})}\text{diag}\left\{\mathbf{T}^{(\mathbf{x})}\right\} \quad , \quad \alpha_x = \boldsymbol{\alpha}_{\boldsymbol{G}}^{(x)}\boldsymbol{o} = \mathbf{P}_{\mathbf{I}}^{(\mathbf{x})}\mathbf{M}_{\mathbf{q}}^{(\mathbf{x})}\mathbf{T}^{(\mathbf{x})} \tag{3.66}$$

Using (3.58), we have,

$$\boldsymbol{\alpha}_{\boldsymbol{G}}^{\Omega} = \mathbf{F}^{\Omega}\text{diag}\left\{\mathbf{A}_{\mathbf{T}}\right\} + \mathbf{F}^{\Omega}\mathbf{A}(\mathbf{e}-\mathbf{B})^{-\mathbf{1}}\text{diag}\left\{\mathbf{B}_{\mathbf{T}}\right\} \ , \ \boldsymbol{\alpha}_{\boldsymbol{G}}^{\Phi} = \mathbf{F}^{\Phi}(\mathbf{e}-\mathbf{B})^{-\mathbf{1}}\text{diag}\left\{\mathbf{B}_{\mathbf{T}}\right\} \tag{3.67}$$

The term $\mathbf{F}^{\Omega}\text{diag}\{\mathbf{A}_{\mathbf{T}}\}$ indicates the probability of making no handoffs and terminating at the initial AGW. We denote this term as $\mathbf{p}_{\mathbf{T_0}}^{\mathbf{G}}$ for the vector form, where each element corresponds of making no handoffs at each initial AGW, and p_{T_0} for making no handoffs in all AGWs (i.e., $p_{T_0} = \mathbf{p}_{\mathbf{T_0}}^{\mathbf{G}}\mathbf{o}$). Notice that $\mathbf{p}_{\mathbf{T_0}}^{\mathbf{G}}$ and p_{T_0} are only defined for on-net sessions, as off-net traffic must make at least one handoff to enter the network under consideration. Thus, $\mathbf{p}_{\mathbf{T_0}}^{\mathbf{G}} = \mathbf{0}$ for offnet traffic. On the other hand, the term $\mathbf{F}^{\Omega}\mathbf{A}(\mathbf{e}-\mathbf{B})^{-\mathbf{1}}\text{diag}\{\mathbf{B}_{\mathbf{T}}\}$ indicates terminating in the network for on-net sessions after making more than one handoff from each AGW. We denote this term as $\mathbf{p}_{T_1}^{\mathbf{G\Omega}}$ in the vector form and $p_{T_1}^{\Omega}$ for the total probability. Clearly, $\mathbf{p}_{T_1}^{G\Phi} = \boldsymbol{\alpha}_{\boldsymbol{G}}^{\Phi}$ and hence $p_{T_1}^{\Phi} = \alpha_\Phi$ for off-net traffic. Notice also that $p_{T_1}^{\Omega} = 1 - p_{T_0} - \beta_\Omega$ and $p_{T_1}^{\Phi} = 1 - \beta_\Phi$. We now proceed to obtain the mean number of interim/reauthentication messages using the obtained values for K_x, β_x, p_{T_0}, and p_{T_1}.

3.6.2.6 The Mean Number of Interim and Reauthentication Messages

In order to solve the expectations in (3.49), we first reformulate it to reflect the types of holding times $i \in \{F, O, Tr, T\}$ as,

$$E\left[\xi_I^{(x)}\right] = \underset{\forall H_{i,j}}{\text{E}}\left[\sum_{n=1}^{N_{H_{i,j}}} I_{H_{i,j}}\right] = \sum_{i\in\{F,O,Tr,T\}}\sum_{j=1}^{N_{AGW}} E\left[N_{H_{i,j}} \,|\, i,j\right] E\left[I_{H_{i,j}} \,|\, i,j\right] \Pr\{i,j\} \tag{3.68}$$

where $\Pr\{AGW = j\}$ is given by $\|\mathbf{F}_{\mathbf{I}}^{(\mathbf{x})}\|_j$ for onnet and offnet sessions and that $E\left[I_{H_{i,j}} \mid i\right]$ is given by (3.48). Due to the complexity of (3.68), let us first simplify our analysis by assuming the same residence time (i.e., $R_j = R$) for all AGWs. We refer to this as *the homogeneous residence time* assumption and relax it at the end of the section. Consequently, there is no more dependence on the serving AGW j, and hence we have $\forall j E\left[I_{H_{i,j}} \mid i\right] = E\left[I_{H_i} \mid i\right]$ and that $N_{H_{i,j}} = N_{H_i}$. It follows that, these terms

factor out of the summation $\sum_{j=1}^{N_{AGW}} \Pr\{i,j\} = \Pr\{i\}$ and (3.68) becomes,

$$E\left[\xi_I^{(x)}\right] = \sum_{i\in\{F,O,Tr,T\}} E\left[I_{H_i} \mid i\right] E\left[N_{H_i} \mid i\right] \Pr\{i\} \tag{3.69}$$

Now the only left parameters to evaluate are $E\left[N_{H_i} \mid i\right]$ and $\Pr\{i\}$. By investigating possible AGW holding time sequences (e.g., the sequence $\{H_O, H_{Tr}, H_{Tr}, H_T\}$), it is clear that full holding times (H_F) occur once if the session makes no handoffs (i.e., $\Pr\{i=F\} = p_{T_0}$ and $E\left[N_{H_F} \mid F\right] = 1$). Similarly, the originating holding times (H_O) occur once if the session makes at least one handoff (i.e., $\Pr\{i=O\} = 1 - p_{T_0}$ and $E\left[N_{H_F} \mid O\right] = 1$). Furthermore, the terminating holding times (H_T) occur if the session makes at least one handoff and does not leave the network (i.e., $\Pr\{i=T\} = p_{T_1}^{(x)}, x \in \{\Omega, \Phi\}$ and $E\left[N_{H_T} \mid T\right] = 1$). Finally, transit holding times (H_{Tr}) occur if the session makes at least two handoffs. Let $N_{HO}^{(x)}$ denote the mean number of handoffs in the network plus the handoffs to roaming partners, then $N_{HO}^{(x)} = k_x + \beta_x$. The number of transit holding times is one less than N_{HO} (e.g., consider a session that makes three handoffs the last was to a roaming partner, we have the sequence H_O,H_{Tr},H_{Tr}). Since $E\left[N_{H_{Tr}} \mid Tr\right] \Pr\{Tr\}$ is equal to $E\left[N_{H_{Tr}} \cap Tr\right]$, we obtain $E\left[N_{H_{Tr}} \cap Tr\right]$ by averaging out over all possible numbers of handoffs as,

$$E\left[N_{H_{Tr}} \cap Tr\right] = \sum_{k=2}^{\infty} (k-1) \Pr\left\{N_{HO}^{(x)} = k\right\} = E[K_x] - p_{T_1}^{(x)} \tag{3.70}$$

Hence the mean number of interim messages $E\left[\xi_I\right]$ under the homogeneous residence time assumption from all AGWs is given as,

$$E\left[\xi_I^{(x)}\right] = p_{T_0} E\left[I_{H_F}\right] + (1-p_{T_0}) E\left[I_{H_O}\right] + \left(E[K_x] - p_{T_1}^{(x)}\right) E\left[I_{H_{Tr}}\right] + p_{T_1}^{(x)} E\left[I_{H_T}\right] \tag{3.71}$$

Now, that we know the form of the solution, let us relax the homogeneous residence time assumption. This means that we need to reconsider the evaluation of the mean number of interims for the four holding time categories for each AGW j (i.e., $E\left[I_{H_{i,j}} \mid i,j\right]$). We also need to consider the joint probability, $\Pr\{i,j\}$, for each holding time category i to occur in AGW j. For originating (O), terminating (T), and full sessions (F), there is always one message independent of the AGW holding time distribution (i.e., $H_{i,j}$). However, their probabilities vary depending on the distribution $H_{i,j}$ and are given by the vector form of the total probabilities used in (3.71). Hence, the probability of occurrence for full sessions is $\Pr\{F,j\} = \|\mathbf{P}_{\mathbf{T_0}}^{\mathbf{G}}\|_j$, for originating sessions is $\Pr\{O,j\} = (1-p_{T_0})\|\mathbf{F}_{\mathbf{I}}^{(\mathbf{x})}\|_j$, and for terminating sessions is $\Pr\{T,j\} = \|\mathbf{P}_{\mathbf{T_1}}^{\mathbf{G}(\mathbf{x})}\|_j$. For transitioning sessions, the expectation $E\left[I_{H_{Tr,j}} \mid Tr,j\right] \Pr\{Tr,j\} = E\left[I_{H_{Tr,j}} \cap Tr,j\right]$,

is given similarly to (3.70). It can be shown that it has the vector form of (3.70), $E\left[I_{H_{Tr,i}} \cap Tr, i\right] = \|\mathbf{K_G^{(x)}} - \mathbf{P_{T_1}^{G(x)}}\|_i$. Hence, the corresponding signaling rate from all AGWs is given as the sum of the interim messages generated by each AGW j, denoted as $\|\boldsymbol{\xi}_I^{(x)}\|_j$, as,

$$\begin{aligned}\|\boldsymbol{\xi}_I^{(x)}\|_j &= \|\mathbf{P_{T_0}^G}\|_j E\left[I_{H_{F,j}}\right] + (1-p_{T_0})\|\mathbf{F_I^{(x)}}\|_j E\left[I_{H_{O,j}}\right] \\ &\quad + \left(\|\mathbf{K_G^{(x)}} - \mathbf{P_{T_1}^{G(x)}}\|_j\right) E\left[I_{H_{Tr,j}}\right] + \|\mathbf{P_{T_1}^{G(x)}}\|_j E\left[I_{H_{T,j}}\right] \\ E\left[\xi_I^{(x)}\right] &= \sum_{j=1}^{N_{\text{AGW}}} \|\boldsymbol{\xi}_I^{(x)}\|_j \end{aligned} \tag{3.72}$$

Now that we have the interims rate from each AGW as in (3.72), the mean number of reauthentications in (3.46) can be obtained similarly to the mean number of interims. However, one should use the authorization lifetime Δ_M instead of the interim interval Δ_T when calculating $E[I_{H_i}]$ in (3.72) and multiply by the number of authentication messages δ depending on whether the home agent request (HAR) is sent (see Fig. 3.11) as,

$$E\left[\xi_{Re}^{(x)}\right] = \delta E\left[\xi_I^{(x)}\right]|_{\Delta_T=\Delta_M} \quad , \quad x \in \{\Omega, \Phi\}$$

3.6.2.7 The Generalized AAA Signaling Rate Model

To evaluate the AAA signaling load in arbitrary AAA deployments, we denote the set of AGWs served by an AAA system (say AAA_n) as $\mathbb{G}_n$. For *centralized* AAA system deployments (e.g., see Fig.3.9(a)), the set $\mathbb{G}_n$ includes all AGWs (i.e, AGW_1-AGW_5). On the other hand, for the *distributed* AAA system deployment (e.g., see Fig.3.9(b)), the set $\mathbb{G}_n$ may not include all AGWs (e.g., for AAA_1 the set $\mathbb{G}_1$ includes AGW_1 and AGW_2, for AAA_2 the set $\mathbb{G}_2$ includes AGW_3-AGW_5, and for AAA_3 the set $\mathbb{G}_3$ includes all AGWs (i.e, AGW_1-AGW_5) for roaming/MVNO users as AAA1 and AAA2 forward roaming sessions towards it). Thus, the signaling rate at any AAA_n is given by the sum of all authentication and accounting messages from its set of AGWs ($\mathbb{G}_n$). Then, substituting (3.47) and (3.72) into (3.46) for on-net (Ω) and off-net (Φ) sessions, we have,

$$E\left[\xi_n\right] = \sum_{x\in\{\Omega,\Phi\}} \sum_{j\in\mathbb{G}_n} \lambda_x \|\mathbf{F_I^{(x)}}\|_j \left[\|\boldsymbol{\nu}_G^{(x)}\|_j (\delta + 2p_a) + p_a\left(\|\boldsymbol{\xi}_I^{(x)}\|_j + \|\boldsymbol{\xi}_{Re}^{(x)}\|_j\right)\right] \tag{3.73}$$

The authentications, accounting start and stop message rates are obtained from (3.47) using (3.61). The rate of accounting interims and reauthentication messages is given by substituting (3.64) and (3.67) into (3.72). It is noteworthy to state that due to the low network departure probability, β_x, the results in Section 3.5 in (3.38) can be used as an upper bound approximation for the AAA signaling estimate in (3.73) under homogeneous residence times (see (3.71)) for short sessions and/or large networks.

3.6.2.8 Approximation for the Generalized AAA Signaling Rate Model

A simplifying approximation for (3.73) can be obtained by assuming exponential residence times (i.e., $k_r^j = 1$ in (3.50)) similar to the approximation used in (3.39). Hence, the AGW holding time distributions become identical (i.e., we drop the index i in (3.72) as $E\left[I_{H_{i,j}}\right] = E\left[I_{H_j}\right]$). Hence, the holding times are redefined using exponential distributions as $f_{H_j}(t) = \frac{1}{E_{H_j}} e^{-\frac{t}{E_{H_j}}}$. The mean holding time E_{H_j} for each AGW is given by dividing the mean session duration E_s by the number of holding times in the session as $E_{H_j} = E_s/(\|\mathbf{P}^{\mathbf{G}}_{\mathbf{T_0}}\|_j + (1-p_{T_0})\|\mathbf{F}^{(\mathbf{x})}_{\mathbf{I}}\|_j + \|\mathbf{K}^{(\mathbf{x})}_{\mathbf{G}} - \mathbf{P}^{\mathbf{G}(\mathbf{x})}_{\mathbf{T_1}}\|_j + \|\mathbf{P}^{\mathbf{G}(\mathbf{x})}_{\mathbf{T_1}}\|_j)$ and hence the session mean $E\left[I_{H_{i,j}}\right]$ in (3.68) can be given as $[e^{\frac{\Delta_T}{E_{H_j}}} - 1]^{-1}$ which is obtained by substituting the exponential distribution for the holding times in (3.48).

3.6.2.9 The Case of Authentication Delegation

Finally, it is noteworthy to state that our approach can also be adapted to estimate the authentication rate in deployment cases where authentication delegation is applied between AAA systems in the network (e.g., hierarchical AAA system designs). For instance, if for roaming users AAA_3 delegated authorization to AAA_1 and AAA_2 in Fig. 3.10(b), then it will only be contacted if mobile users cross from AGWs served by AAA_1 to AGWs served by AAA_2 (i.e., AGW2 to AGW3 in our example). The authentication rates observed at AAA_1 and AAA_2 are calculated using (3.47), however, the rate at AAA_3 is no longer the sum of both rates for roaming users and is upper limited[7] by the mean crossing rate from AGW_2 to AGW_3. Let $\mathbb{G}_1$ denote the set of AGW1 and AGW2 and the set $\mathbb{G}_2$ denote the set of AGWs 3-5, then, it follows that the authentication signaling rate observed at AAA_3 is given by $\lambda_x \left(\|\mathbf{Q}^{(\mathbf{x})}\|_{2,3} \|\mathbf{P}^{(\mathbf{x})}_{\mathbf{I}} \mathbf{M}^{(\mathbf{x})}_{\mathbf{q}} \mathbf{D}^{(\mathbf{x})}\|_2 + \|\mathbf{Q}^{(\mathbf{x})}\|_{3,2} \|\mathbf{P}^{(\mathbf{x})}_{\mathbf{I}} \mathbf{M}^{(\mathbf{x})}_{\mathbf{q}} \mathbf{D}^{(\mathbf{x})}\|_3 \right)$. The term $\|\mathbf{Q}^{(\mathbf{x})}\|_{j,k} \|\mathbf{P}^{(\mathbf{x})}_{\mathbf{I}} \mathbf{M}^{(\mathbf{x})}_{\mathbf{q}} \mathbf{D}^{(\mathbf{x})}\|_j$ represents the proportion of the number of visits from AGW_j to AGW_k.

3.6.3 Model's Limitations

In addition to some of the limitations for the fixed rate model relevant to the session arrival process, the processing power, and the users' quota, the following are the limitations of the generalized AAA model in (3.73),

[7] Due to possible caching on AAA1 and AAA2, AAA_3 may not always be contacted for each crossing between AGW2 and AGW3. For brevity, we only investigate upper limits.

1. *The exponential session time*: This assumption affects the AGW holding time predictions H_i as discussed in Section 3.5.5. Unlike the mean number of handoffs at the home AAA which was derived in Section 3.5 in (3.37), the mean number of handoffs observed in a distributed AAA system or when roaming depends on the session distribution. We resolve this aspect in Section 3.8.1 by assuming a generic session distribution for obtaining the mean number of handoffs. However, the holding time distributions at the k^{th} handoff remain open for further research.

2. *The mobility model*: In our formulation, we have assumed random mobility between AGWs. However, in reality other movement patterns such as straight movements or more correlated patterns may be incurred. This limitation primarily affects the calculation of the mean number of handoffs between AGWs and the incurred roaming liklihoods. We resolve this issue in Section 3.8.2.

3.7 Applications in Today's AA Schemes

In this section, we discuss the use of the proposed models in this chapter to evaluate the signaling load due to known Authentication and Authorization (AA) signaling optimization methods. In addition, we shortly discuss the impact of context transfer between AGWs on the resulting AAA signaling load.

3.7.1 Authentication Signaling for Wireless Network Association

While we have only focused on the authentication due to the movement between AGWs and due to Mobile IP signaling [61], we show that only straightforward changes are needed to accommodate AA signaling for wireless network association. This type of authentication is needed to secure the airlink traffic and varies in terms of the number of signaling messages and their sizes depending on the used cellular technology. For instance, 1xEVDO [56] uses one exchange with the AAA infrastructure over the so-called A12 authentication based on CHAP. However, WiMAX and LTE systems adopt EAP based authentication schemes (e.g., EAP-TLS, EAP-TTLS). EAP methods usually entail several exchanges with the AAA system. As shown in Fig. 3.12, EAP signaling involves N exchanges (e.g., 12 messages for EAP-TTLS) between the mobile node and the AAA framework to establish a master session key. The master key is then transferred to the AGW which in turn derives authentication keys and conveys them to the serving base station. Afterwards, the mobile node establishes security associations (SA) and traffic encryption keys (TEK) with the serving base station using five messages [122].

In mobile networks, EAP authentication can result in large delays, signaling overhead, and air link loading due to AGW handoffs. Hence, optimizations have been proposed to minimize the authentication signaling after the first N exchanges [31, 33]. This is

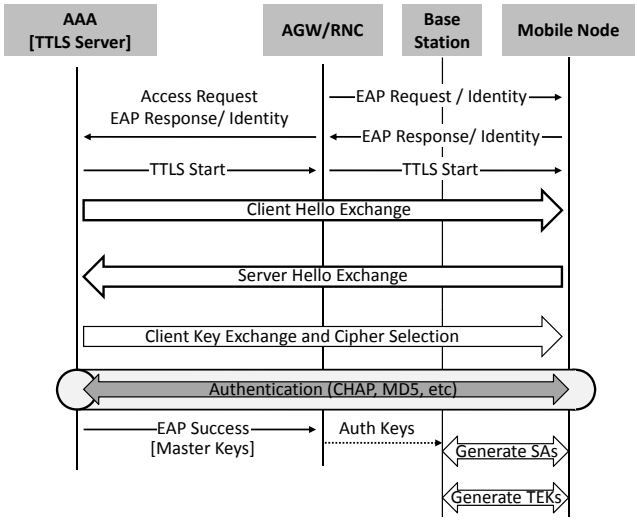


Figure 3.12: A simplified EAP-TTLS signaling flow [SA: Security Association, TEK: Traffic Encryption Key].

primarily achieved by (i) delegating further authentications to the visited AAA system and avoiding contacting the home AAA [32, 33], and (ii) by modifying EAP signaling to finish in one exchange with the AAA system during handoffs [31]. The modified EAP signaling can largely reduce the incurred signaling delay and air link load pertaining to authentications and can be effectively combined with authentication delegation to further reduce the signaling load between visited and home networks in roaming scenarios.

3.7.1.1 The Air Link Load

In order to obtain the consumed wireless link capacity due to authentication signaling for network association, let us first discuss how the Total Air link Load (TAL) per authentication operation can be obtained. To this end, we adopt a similar approach as in [110] and assume the use of radio link protocol acknowledgements to ensure better reliability for authentication signaling. Let us denote the authentication message size as L_y where y denotes the authentication scheme (e.g., A12 or EAP). The number of radio frames to transport L_y is given as $n = \lceil \frac{L_y}{L_R} \rceil$ where L_R is the radio frame size in bits. Assuming a typical three retransmissions for the Radio Link Protocol (RLP) due to erroneous frames and given the radio link frame error rate e, the transmission overhead

of the radio frames including retransmissions is given as [110],

$$O_R = L_R(1-e) + \sum_{i=1}^{3}\sum_{j=1}^{i} P(C_{ij})(2iL_R)\,,\; P(C_{ij}) = e(1-e)^2\left((2-e)e\right)\left(\frac{i^2-i}{2}+j-1\right)$$

Hence the total overhead per authentication message is simply given by the product of the number of the radio frames and the overhead per radio frame as,

$$O_y = nO_R$$

If the packet loss probability in the wired network is p_w, then according to [110], the overall packet loss probability due to the wired and wireless network losses is given as,

$$p = 1 - \left[1 - e\left((2-e)e\right)^6\right]^n (1-p_w)$$

Assuming m retransmissions for authentication packets, the total airlink load per packet including retransmissions can be written as the product of the mean number of retransmissions per frame and the number of frames per authentication message as,

$$\mu_y = \sum_{i=1}^{m} \frac{p^{i-1}(1-p)}{1-p^{m+1}}(iO_y) = O_y\kappa \quad , \quad \kappa = \left(\frac{1}{1-p} - \frac{(m+1)\,p^{m+1}}{1-p^{m+1}}\right)$$

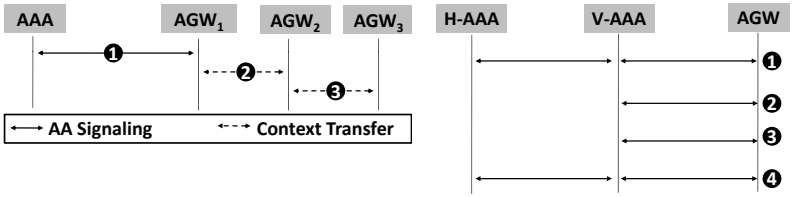
Since each authentication operation may incur multiple (i.e., η_y) exchanges over the airlink (e.g., $\eta_{\text{EAP-TTLS}}$[8] $= 17$), the Total Airlink Load per authentication operation (TAL) in bits is given by the sum of the overhead from all messages as,

$$T = \sum_{i=1}^{\eta_y} \mu_{y_i}$$

where μ_{y_i} is the message size of the i^{th} authentication message. When authentication optimization methods such as in [31] are used, the TAL for initial authentications and reauthentications (denoted as T_0) maybe different from that following AGW handoffs (denoted as T_1). This is because only one exchange is needed with the AAA system after handoffs, while for session initiation twelve messages are used in EAP-TTLS.

Now that we have estimated the TAL per authentication operation, we use the model in (3.38) to estimate the consumed airlink capacity (in bps) from all users per base station. Let N_c be the total number of cells in an AGW area and N_b be the number of cells in

[8] 12 EAP messages with the AAA system plus 5 key establishment negotiation messages with the base station



(a) Context transfer between AGWs (b) AA delegation between AAAs

Figure 3.13: The concept of context transfers and the authentication delegation

the boundary between neighboring AGWs. The interior cells incur authentications at the beginning of the session from on-net traffic and reauthentications from both on-net and off-net sessions. The boundary cells, however, incur extra load due to the handoff sessions. The session arrival rate per cell is $\frac{\lambda \|\mathbf{F}_{\mathbf{I}}^{(\mathbf{x})}\|_i}{N_c}$. Recall that $\frac{E_s}{E_R}$ is the mean number of handoffs. Assuming i.i.d residence times for AGWs and that the load is uniformly distributed in the network, the used airlink capacities for interior cells ς_I and boundary cells ς_B are given as,

$$\varsigma_I = \frac{\lambda T_0 \left(1 + E\left[\xi_{Re}\right]\right)}{N_c N_{AGW}} = \frac{\lambda T_0 \left(1 + p_a E\left[M\right]\right)}{N_c N_{AGW}} \quad , \quad \varsigma_B = \varsigma_I + \frac{\lambda E_s T_1}{E_R N_b N_{AGW}} \tag{3.74}$$

This result can be generalized as in Section 3.6 for non-uniform load and AGW residence times using the specific number of messages per AGW (j) to reflect the load at each AGW for on-net and off-net sessions as,

$$\varsigma_{I_j} = \sum_{x \in \{\Omega, \Phi\}} \frac{\lambda_x T_0 \left(\|\mathbf{F}_{\mathbf{I}}^{(\mathbf{x})}\|_j + p_a \|\boldsymbol{\xi}_{Re}^{(x)}\|_j \right)}{N_c} \quad , \varsigma_{B_j} = \sum_{x \in \{\Omega, \Phi\}} \varsigma_{I_j} + \frac{\lambda_x \|\mathbf{K}_{\mathbf{G}}^{(\mathbf{x})}\|_j T_1}{N_b} \tag{3.75}$$

3.7.2 Context Transfers and Authentication Delegation

Context transfers between AGWs were proposed to minimize the time a new AGW needs to authorize handoff sessions such as the CXTP protocol in [123]. The use of such mechanisms allows the transfer of session context information between AGWs when the mobile user crosses the boundary between two AGWs, and hence the authentication with the target AGW upon handoffs is no longer necessary. Depending on how reauthentications are triggered[9] (i.e., based on the session initiation time or the

[9]This aspect is not currently standardized so we discuss all possibilities.

latest handoff instant), the number of authentication and reauthentications is obtained either using the fixed model in (3.11) for the first case, or otherwise using (3.31) as $\sum_{i=1}^{N_{\text{AGWs}}} \lambda^{(i)} (1 + p_a E[M])$. Since context transfer signaling takes place in the core IP network, the effect of signaling packet losses is insignificant.

Relevant to EAP signaling, authentication delegation can highly reduce the signaling load on the visited and home operators' networks. It can also be used to speed up the authentication process especially in roaming scenarios by delegating the authentication to the (visited) V-AAA system located in the visited network [32, 33]. In such cases, the (home) H-AAA is only contacted during initial session authorization in order to obtain the user's profile as shown in Fig. 3.13(b). Afterwards, the V-AAA authorizes authentication requests due to AGW handoffs (steps 2-3 in Fig. 3.13(b)). When deemed necessary such as when the delegation period expires, the V-AAA may contact the H-AAA for authentication (step 4). Clearly, delegation not only reduces the authentication time but also reduces the load on the link between the V-AAA and the H-AAA by eliminating the need to proxy AA requests to the H-AAA.

Table 3.3: The signaling load per session for various authentication mechanisms [In EAP based schemes, the extra 5 messages are used for security association and traffic encryption key negotiations between the mobile node and the serving base station]

Auth Mechanism	**No. Exchanges V-AAA**	**Proxy Operations**	**Airlink Overhead**
Basic MobileIP	$(1+E[K_x])+p_aE[\xi_{Re}^{(x)}]$	equals V-AAA	$T_0 = \eta_{MIPReg}$, $T_1 = \eta_{MIPReg}$
Basic MobileIP with delegation	$(1+E[K_x])+p_aE[\xi_{Re}^{(x)}]$	one exchange	$T_0 = \eta_{MIPReg}$, $T_1 = \eta_{MIPReg}$
Context transfer with delegation	$1+p_aE[\xi_{Re}^{(x)}]$	one exchange	$T_0 = \eta_{MIPReg}$, $T_1 = \eta_{MIPReg}$
EAP	$N(1+E[K_x])+Np_aE[\xi_{Re}^{(x)}]$	same as V-AAA	$T_0 = \sum_{i=1}^{N+5} \eta_{EAP_i}$, $T_1 = \sum_{i=1}^{N+5} \eta_{EAP_i}$
EAP with delegation	$N(1+E[K_x])+p_aNE[\xi_{Re}^{(x)}]$	N exchanges	$T_0 = \sum_{i=1}^{N+5} \eta_{EAP_i}$, $T_1 = \sum_{i=1}^{N+5} \eta_{EAP_i}$
EAP with optimization [31]	$N+E[K_x]+p_aNE[\xi_{Re}^{(x)}]$	same as V-AAA	$T_0 = \sum_{i=1}^{N+5} \eta_{EAP_i}$, $T_1 = \sum_{i=1}^{6} \eta_{EAP_i}$
EAP with opt. delegation	$N+E[K_x]+p_aNE[\xi_{Re}^{(x)}]$	N exchanges	$T_0 = \sum_{i=1}^{N+5} \eta_{EAP_i}$, $T_1 = \sum_{i=1}^{6} \eta_{EAP_i}$
EVDO A12 Auth.	$(1+E[K_x])+p_aE[\xi_{Re}^{(x)}]$	same as V-AAA	$T_0 = \eta_{A12}$, $T_1 = \eta_{A12}$
EVDO A12 Auth. with delegation	$(1+E[K_x])+p_aE[\xi_{Re}^{(x)}]$	one exchange	$T_0 = \eta_{A12}$, $T_1 = \eta_{A12}$

3.7.2.1 Summary of AA Methods

In this section, we provide comparison models for the signaling load and the airlink overhead based on the models developed in this chapter. We used the generalized AAA model in (3.73) as basis for our notation. Notice that adapting the fixed model in (3.11) and the mobile models in (3.38) is straightforward. For the fixed model, $E[K_x] = 0$ and the sub/superscript x is dropped as sessions are always on-net. For the mobile model, x is dropped and $E[K_x] = \frac{E_s}{E_r}$. In both cases, the V-AAA to H-AAA signaling is generally irrelevant by definition. Table 3.3 summarizes our models.

We first include the signaling load and the airlink overhead due to MobileIP authentication signaling, shown in Fig. 3.4. We show the effect of possible optimizations including context transfers and authorization delegation. Since the airlink only carries MobileIP registration messages, the message size parameters $T_0 = T_1$ only reflect the size of such messages. However, for EAP methods the parameters T_0 and T_1 reflect the sizes of EAP messages before and after handoffs which may differ depending on whether handoff optimization techniques (e.g., [31]) are implemented. Finally, for EVDO systems, the authentication process only involves one message and hence T_0 and T_1 reflect the size of that message.

3.8 Towards Generalized Handoff Modeling

The main goal of this section is to develop a theoretical framework to estimate the mean number of handoffs under generic assumptions of session duration, mobility pattern between AGWs, number of cells per AGW, and users' distribution. In Section 3.8.1, we show how to relax the exponential session assumption for a simple random mobility pattern in one dimension. In Section 3.8.2, we our results to any arbitrary Markovian mobility model. In Section 3.8.3, we derive the AGW residence time based on the cellular residence times. We conclude this section by outlining a generic hierarchical scheme where the mean number of handoffs is derived depending on arbitrary number of cells per AGW and mobility patterns between them (see Section 3.8.4).

3.8.1 Generalizing the Session Statistics

3.8.1.1 Assumptions

1. The arrival processes for on-net and off-net traffic are Poissonian with mean rates of λ_Ω and λ_Φ respectively (as in Section 3.6.2.1).

2. The session duration, S, and residence times, R, are independent and identically distributed (i.i.d) with generic distributions and existing Laplace transforms.

3. The network contains N_{AGW} linearly arranged AGWs indexed from 0 to $N_{\text{AGW}} - 1$ with two large roaming partners at the edges.

4. The users' mobility behavior between gateways is *random*.

3.8.1.2 Analysis

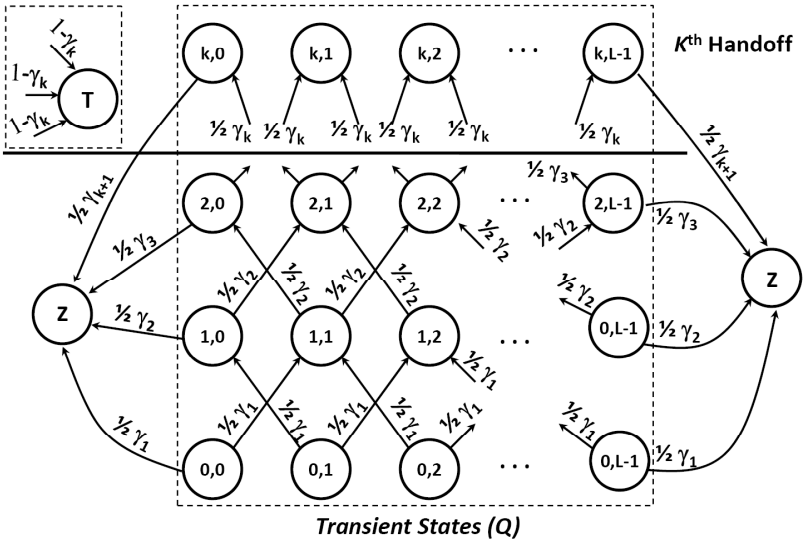


Figure 3.14: Markovian mobility model under generalized session assumptions (adapted from [124]) [L denotes the number of AGWs (N_{AGW}) and is used for the clarity of the figure. All transient states can reach state T, state Z is drawn twice for clarity].

In order to evaluate the mean number of handoffs for on-net and off-net sessions[10] ($E[N_x]$, $x \in \{\Omega, \Phi\}$), we extend the transient Markov chain analysis in Section 3.6 for general session durations. This is achieved by tracking the history of the completed handoffs when calculating the transition probabilities. This is actually the trick we use to relax the exponential session time which intrinsically 'forgets' the handoff history. Similar to the model in Section 3.6, the state description only depends the number of AGWs in the network and is unaffected by neither the session nor the residence time distributions.

As shown in Fig. 3.14, our transient chain consists of a 2-tuple transient state definition

[10] See Section 3.6.2 for the definition of on-net and off-net sessions.

$\delta_j^k = (k, j)$ representing a session served by the j^{th} AGW after completing k handoffs, and two absorbing states: Z representing a session leaving the domain to *any* of the roaming partners and T representing the session termination. The transition probability is given by $0.5\gamma_k^x$ where γ_k^x is defined as the probability that the session contains at least one more handoff given that it had already made $k-1$ handoffs. Thus, after a residence time R elapses, the user either moves east or west with a probability of $0.5\gamma_k^x$ or if the session has terminated, the chain goes to the absorbing state T with a probability of $1-\gamma_k^x$. If the session is in states δ_0^k or $\delta_{N_{\text{AGW}}-1}^k$, then a user may move out to the neighboring areas (i.e., state Z) with a probability of $.5\gamma_k^x$. Before proceeding, it is noteworthy to emphasize that we use 2-tuple state definitions as we have generic session and residence times. Thus, the transition probabilities are only a function of the current transient state which provides full information about the possible next AGWs and the total number of completed handoffs.

Markov Chain Formulation

Let us define $\mathbf{Q}^{(\mathbf{x})}$ as the transition probability matrix among transient states and $\mathbf{U}_{\mathbf{Z}}^{(\mathbf{x})}$ as the probability vector of moving to the absorbing state Z for both on-net and off-net arrivals. $\mathbf{Q}^{(\mathbf{x})}$ is defined as a $\kappa N_{\text{AGW}} \times \kappa N_{\text{AGW}}$ matrix with $\kappa \to \infty$ while $\mathbf{U}_{\mathbf{Z}}^{(\mathbf{x})}$ is a $KN_{\text{AGW}} \times 1$ vector. κ goes to infinity as it is possible that a session makes infinite number of handoffs in a domain. Ordering our states lexographically as $(0,0),(0,1),...,(1,0),...$, the transition probabilities among transient states $\mathbf{Q}^{(\mathbf{x})}$ (i.e., representing handoffs between AGWs in the network) and to the absorbing states, $\mathbf{U}_{\mathbf{z}}^{(\mathbf{x})}$, (i.e., representing roaming to other networks) are given as,

$$\mathbf{Q}^{(\mathbf{x})} = \begin{bmatrix} 0 & \mathbf{D_0} & 0 & 0 & \dots \\ 0 & 0 & \mathbf{D_1} & 0 & 0 \\ 0 & 0 & 0 & \mathbf{D_2} & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix} \quad , \quad \mathbf{U}_{\mathbf{z}}^{(\mathbf{x})} = \begin{bmatrix} \mathbf{C_0} \\ \mathbf{C_1} \\ \mathbf{C_2} \\ \vdots \end{bmatrix} \tag{3.76}$$

where $\mathbf{D_k}$ is an $N_{\text{AGW}} \times N_{\text{AGW}}$ matrix and $\mathbf{C_k}$ is a $N_{\text{AGW}} \times 1$ vector as the product of a mobility matrix component and the chance of making one more handoff as,

$$\mathbf{D_k} = \gamma_{(k+1)}^x \mathbf{Q_M} \quad , \quad \mathbf{C_k} = \gamma_{(k+1)}^x \mathbf{A_M} \tag{3.77}$$

where the mobility matrices inside the network, $\mathbf{Q_M}$, and to the roaming partners, $\mathbf{A_M}$, are given as,

$$\mathbf{Q_M} = \begin{bmatrix} 0 & .5 & 0 & \dots & \\ .5 & 0 & .5 & 0 & \dots \\ 0 & .5 & 0 & .5 & \dots \\ 0 & \vdots & \vdots & & \dots \\ 0 & 0 & \dots & .5 & 0 \end{bmatrix} \quad , \quad \mathbf{A_M} = \begin{bmatrix} .5 \\ 0 \\ 0 \\ \vdots \\ .5 \end{bmatrix} \tag{3.78}$$

Solving the Chain

As we discussed in Section 3.6.2.4, the mean number of handoffs, $E[K_x]$, can be viewed

as the mean number of state *re-visits (i.e., excluding the initial arrival)* among the transient states before absorption (i.e., termination or roaming). The probability of roaming, β_x, is characterized by the last transition to state Z. To find such means, we use the probabilities $\mathbf{F}^{(\mathbf{\Omega})}$ and $\mathbf{F}^{(\mathbf{\Phi})}$ which are defined similarly to Section 3.6.2.4 and denote the likelihoods of initiating sessions from within each AGW. Thus, the initial state probabilities $\mathbf{P}_{\mathbf{I}}^{(\mathbf{x})}$ are given in terms of $\mathbf{F}^{(\mathbf{x})}$ and have sizes of $1 \times \kappa N_{\mathrm{AGW}}$, $\kappa \to \infty$, as,

$$\begin{aligned}\mathbf{P}_{\mathbf{I}}^{(\mathbf{\Omega})} &= [f_{\Omega}(0), \ldots, f_{\Omega}(N_{\mathrm{AGW}}-1)\,0\ldots] \\ \mathbf{P}_{\mathbf{I}}^{(\mathbf{\Phi})} &= [f_{\Phi}(0), 0, \ldots, f_{\Phi}(N_{\mathrm{AGW}}-1)\,0\ldots]\end{aligned} \tag{3.79}$$

Notice that the initial states are restricted to the $k=0$ states (i.e., the bottom level). This assumption is obvious for on-net traffic while for off-net arrivals, however, the history of the session is generally unknown because the residence time statistics and the size of the roaming, often competing, partners' networks are typically not available. In this case (i.e., off-net arrivals), we exploit the residual of the service session duration, denoted as $\hat{S}$, and, consequently, we only track the handoff history in the domain under analysis(i.e., by starting from the $k=0$ states).

Denoting the conditional mean of revisits to each state, δ_i^k, given the initial state δ_j^0 as $E\left[K_x(k,i) \mid \delta_j^0\right]$, then using the initial probabilities in (3.79), the mean number of handoffs inside the network is given as,

$$E[K_x] = \sum_{j=0}^{N_{\mathrm{AGW}}-1} \sum_{k=0}^{\kappa-1} \sum_{i=0}^{N_{\mathrm{AGW}}-1} E\left[K_x(k,i) \mid \delta_j^0\right] f_x(j) \tag{3.80}$$

Similarly, denoting the probability that a user roams given the initial state δ_j^0 as $\beta_x(j)$, then the roaming probability is given as,

$$\beta_x = \sum_{j=0}^{N_{\mathrm{AGW}}-1} \beta_x(j) f_x(j) \tag{3.81}$$

From the literature [120], the mean number of *visits (i.e., including the first arrival)* before absorption to a given transient state δ_i^k given the initial state, δ_j^0, is expressed by the elements of the fundamental matrix $\mathbf{M}_{\mathbf{z}}^{(\mathbf{x})} = \left(\mathbf{e} - \mathbf{Q}^{(\mathbf{x})}\right)^{-1}$ where $\mathbf{e}$ is the identity matrix. It follows that the sum of the elements in the j^{th} row is equal to

$$\sum_{\forall \mathrm{cols} \in \mathrm{row}\ j} \|\mathbf{M}_{\mathbf{z}}^{(\mathbf{x})}\| = \sum_{k=0}^{\kappa-1} \sum_{i=0}^{N_{\mathrm{AGW}}-1} E\left[K_x^{(k,i)} \mid \delta_j^0\right] + 1 \tag{3.82}$$

We now solve for the mean number of handoffs, $E[K_x]$, and the roaming probability, β_x by using the results from the transient Markov chains theory in [120]. Using the initial state probabilities in (3.79) and substituting (3.82) into (3.80), it follows that the mean number of handoffs in the network, $E[K_x]$, can be rewritten as,

$$E[K_x] = -1 + \sum_{m=0}^{N_{\text{AGW}}-1} \sum_{n=0}^{\kappa-1} f_x(m) \, \| \mathbf{M}_{\mathbf{z}}^{(x)}_{(m,n)} \| \tag{3.83}$$

Furthermore, the roaming probability, β_x, is given as,

$$\beta_x = \sum_{m=0}^{N_{\text{AGW}}-1} \sum_{n=0}^{\kappa-1} f_x(m) \, \beta_x(j) = \sum_{m=0}^{N_{\text{AGW}}-1} \sum_{n=0}^{\kappa-1} f_x(m) \, \mathbf{M}_{\mathbf{z}}^{(\mathbf{x})} \mathbf{U}_{\mathbf{z}}^{(\mathbf{x})} \tag{3.84}$$

To use (3.83)-(3.84), however, we truncate the state space in Fig. 3.14 to a suitable value of κ (i.e., corresponding to an unlikely number of handoffs, $\Pr(K_x = \kappa) \approx 0$), and hence minimal computational error. $\Pr(K_x = \kappa)$ can be evaluated using results from [17]. A key advantage to our formulation in (3.83) and (3.84) is that $\mathbf{e} - \mathbf{Q}^{(\mathbf{x})}$ has an upper diagonal structure, which allows us to use a *simple backward substitution instead of matrix inversion* to obtain $\mathbf{M}_{\mathbf{z}}^{(\mathbf{x})}$ leading to significant gain in computational efficiency. In Section 3.8.2, we use complex variable analysis and show how to obtain a closed form solution for (3.83)-(3.84). We now proceed to illustrate how the transition probabilities γ_k^x can be calculated.

Obtaining the Transition Probabilities

For on-net arrivals, the instants when the session starts and the user enters an AGW region do not necessarily occur at the same time (i.e., the first handoff happens if *the session duration exceeds the residual of the residence time* $\tilde{R}$), hence the k^{th} handoff happens if the session S is longer than the sum of the accumulated residence times as $\tilde{R} + \sum_{i=2}^{k} R \leq S$. On the other hand, for off-net traffic, since it always enters with already established sessions, we assume that the first handoff in the domain occurs if the *residual of the session time,* $\tilde{S}$, exceeds the *residence time* R. It follows that the k^{th} handoff happens if $\sum_{i=1}^{k} R \leq \tilde{S}$.

Defining $f_R^*(\hat{s}), f_{\tilde{R}}^*(s), f_S^*(\hat{s})$, and $f_{\tilde{S}}^*(\hat{s})$ as the Laplace transforms of $R, \tilde{R}, S$, and $\tilde{S}$, using the results from [17] for on-net arrivals, and extending the results to incorporate off-net traffic using the residual of the session duration, γ_k^x can be obtained as,

$$\gamma_k^x = \begin{cases} \dfrac{\Pr\left\{\tilde{R} + \sum_{i=2}^{k} R \leq S\right\}}{\Pr\left\{\tilde{R} + \sum_{i=2}^{k-1} R \leq S\right\}} = \dfrac{G_\Omega(k)}{G_\Omega(k-1)} & x = \Omega \\ \dfrac{\Pr\left\{\sum_{i=1}^{k} R \leq \tilde{S}\right\}}{\Pr\left\{\sum_{i=1}^{k-1} R \leq \hat{S}\right\}} = \dfrac{G_\Phi(k)}{G_\Phi(k-1)} & x = \Phi \end{cases} \tag{3.85}$$

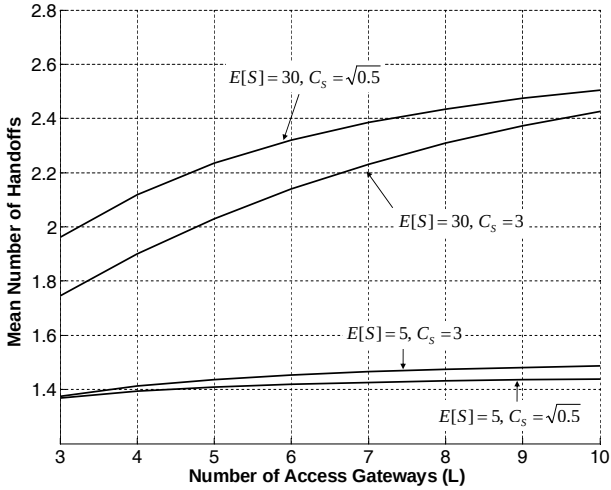


Figure 3.15: Mean number of handoffs as a function of the number of AGWs (L), (adapted from [124]) [R follows Gamma distribution with $E[R] = 20$ mins, $c = 3$].

where $G_x(k)$ is obtained using the residues at the poles p_i in the right half plane from $f_S(-\hat{s})$ (see [17] for examples) as,

$$G_x(k) = \begin{cases} -\sum_i \underset{\hat{s}=p_i}{\text{Res}} \dfrac{f_{\tilde{R}}^*(\hat{s})\left(f_R^*(\hat{s})\right)^{k-1} f_S(-\hat{s})}{\hat{s}} & x = \Omega \\ -\sum_i \underset{\hat{s}=p_i}{\text{Res}} \dfrac{\left(f_R^*(\hat{s})\right)^k}{\hat{s}^2 E[S]} f_S(-\hat{s}) & x = \Phi \end{cases} \tag{3.86}$$

More details on (3.85) is available in our work in [107, 108].

Numerical Example
To illustrate our model, we study the handoff signaling rate (i.e., $\sum_x \lambda_x (E[K_x] + \beta_x)$) for four exemplary services (e.g., VoIP and video), with all combinations of mean durations of 5 and 30 mins and coefficients of variation, c, of $\sqrt{0.5}$ (i.e., Erlang) and 3 (i.e., Hyper-exponential). The initial arrivals are uniformly distributed throughout all AGWs for on-net traffic and are evenly balanced for off-net traffic. The arrival rates are $\lambda_\Omega = 1$ req/s and $\lambda_\Phi = 0.1$ req/s and are fixed for comparison purposes. For all cases, we vary the number of AGWs from 2 to 10 and calculate the corresponding handoff signaling rate. As shown in Fig.3.15, we observe the following: 1) The signaling rate is a non-linear and non-decreasing function of the number of AGWs for all services even at fixed

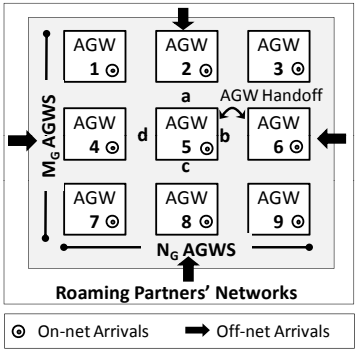


Figure 3.16: Sample topology with $n = 9$ Access Gateways (AGW), (adapted from [125]) [The borders of every gateway are marked a,b,c,d corresponding to north, east, south and west movements (e.g., AGW 5)].

arrival rates, λ_x. 2) The signaling rate from services with relatively low means (i.e., to the residence time such as cases A and B in Figure 3.15) are less affected by the increase in the number of AGWs as they are unlikely to make many handoffs and hence terminate within the network. 3) The effect of the coefficient of variation is more pronounced on services with longer durations (i.e., cases C, D). After making these conclusions, let us now proceed to generalize the mobility pattern from linear random movement (i.e., left and right) to more generic patterns and see how we can avoid the matrix truncation process in (3.82).

3.8.2 Generalizing the Mobility Model

The areas a mobile may traverse during a session can have any arbitrary arrangement and the sessions may move in arbitrary patterns. In this section, we show how to relax the random movement assumption between AGWs. To simplify the discussion, let us first consider an area consisting of rectangularly arranged AGWs (i.e., $M_g \times N_g$ number of gateways within the network under consideration as shown in Fig. 3.16. *We emphasize that the method that we discuss directly applies to any arbitrary topology and is not restricted to regular arrangements*.

When a mobile enters the network coverage area, its future movement is described by a set of transition probabilities which depend on the entering and exit borders [126]. Thus each AGW is described by 4×4 transition probabilities. Using the border labeling shown in Fig. 3.16 for AGW 5, the transition probabilities are denoted as p_{jrl}, where j denotes the AGW, r and l define the entering and the exit borders respectively such that $(r,l) \in \{a,b,c,d\}$. The transition probabilities can be arranged into a set of two

different one step transition matrices, one that defines the initial transition probabilities $\mathbf{P}_{\mathbf{MI}}^{(\mathbf{x})}$, $x \in \{\Omega, \Phi\}$ (i.e., when the session is first served by the network) and another, $\mathbf{P_M}$ that defines the transition probabilities afterwards (i.e., after the initial handoff) as,

$$\mathbf{P}_{\mathbf{MI}}^{(\mathbf{x})} = \begin{pmatrix} \mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})} & \mathbf{A}_{\mathbf{MI}}^{(\mathbf{x})} \end{pmatrix} \quad , \quad \mathbf{P_M} = \begin{pmatrix} \mathbf{Q_M} & \mathbf{A_M} \end{pmatrix} \tag{3.87}$$

Assuming a network of $N_{\text{AGW}} = M_g \times N_g$ access gateways, the matrix $\mathbf{Q_M}$ describes the movement of a handoff session between AGWs and has $4N_{\text{AGW}} \times 4N_{\text{AGW}}$ elements describing movements between neighboring AGWs. For example, a session leaving AGW 5 in Fig. 3.16 at border a enters AGW 2 at border c. Thus, the exit border a of AGW j is linked to the entry border c of AGW $(j - N_g)$, where N_g is the number of AGWs in a row. Let us number the columns and rows of $\mathbf{Q_M}$ by the entry borders of the AGWs as $1a, 1b, 1c, 1d, 2a, 2b, 2c, 2d, ..., N_{\text{AGW}}a, N_{\text{AGW}}b, N_{\text{AGW}}c, N_{\text{AGW}}d$. For a given entry border there are up to four possible transitions. For instance, entering from border a in the j^{th} AGW (i.e. $(ja)^{th}$ row of $\mathbf{Q_M}$), the transition probabilities (which sum to one) are $p_{jaa}, p_{jab}, p_{jac}, p_{jad}$. They have the column numbers $(j - N_g)\,c$, $(j+1)\,d$, $(j + N_g)\,a$ and $(j-1)\,b$, respectively, [126]. Additionally for an AGW at the network boundary, all transitions to the roaming partners (Z) are listed in the $1 \times 4N_{\text{AGW}}$ matrix $\mathbf{A_M}$. An example for AGW 2 (rows $(2a)$ and $(2b)$) is shown for the matrixes $\mathbf{Q_M}$ and $\mathbf{A_M}$ below, where the state numbering is added on top and to the right side for clarity.

$$\mathbf{Q}_M = \begin{array}{c} \begin{matrix} (1a) & (1b) & \cdots & (3d) & \cdots & (5a) & \cdots & & & (\mathbf{Z}) \end{matrix} \\ \left(\begin{array}{ccccccc|c} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \\ 0 & p_{2ad} & \cdots & p_{2ab} & \cdots & p_{2ac} & \cdots & (2a) & & \\ 0 & p_{2bd} & \cdots & p_{2bb} & \cdots & p_{2bc} & \cdots & (2b) & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \end{array}\right) \end{array}, \quad \mathbf{A}_m = \begin{pmatrix} \vdots \\ p_{2aa} \\ p_{2ba} \\ \vdots \end{pmatrix}$$

For on-net sessions, the matrix $\mathbf{Q_{MI}}$ is of dimension $N_{\text{AGW}} \times 4N_{\text{AGW}}$ and contains the transition probabilities for a new session. A session starting in AGW j and leaving at border y corresponds to the j^{th} row of $\mathbf{Q_{MI}}$ and has transition probabilities $\hat{p}_{ja}, \hat{p}_{jb}, \hat{p}_{jc}, \hat{p}_{jd}$. Again, all transitions to the roaming partners created by boundary AGWs are combined in the matrix $\mathbf{A_{MI}}$. For off-net sessions, the matricides $\mathbf{Q_{MI}}$ and $\mathbf{A_{MI}}$ are equal to $\mathbf{Q_M}$ and $\mathbf{A_M}$ respectively. This is because sessions start around and leave the borders of AGWs. Finally, it should be noted that our analysis is independent of the grid like arrangement of AGWs. For other AGW arrangements (e.g., irregular arrangements of AGW areas with more than 4 neighbors), we simply add rows and columns with the corresponding transition probabilities for each edge in the mobility matrices in (3.87).

The Mean Number of Handoffs

In this subsection, we derive the mean number of handoffs, $E\{K_x\}$, for sessions partially served by a network comprised of $M_g \times N_g$ AGWs by extending the transient Markov model in Section 3.8.1 to incorporate the pixel mobility model.

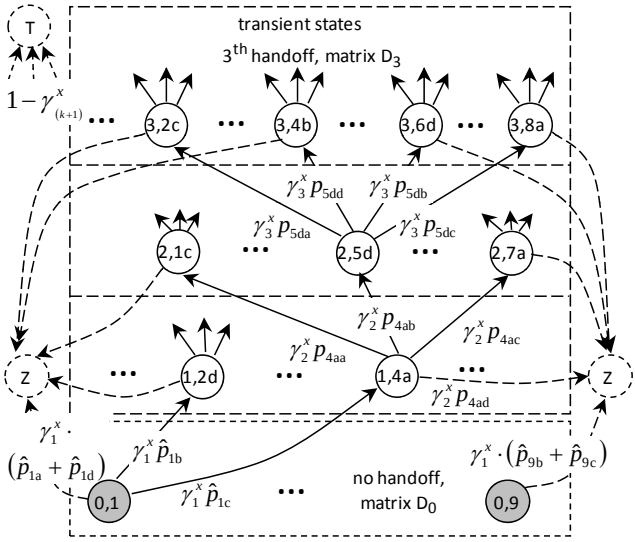


Figure 3.17: Model diagram for the network of Figure 3.16 with selected states and transitions only (adapted from [125]) [all transient states can reach state T, state Z is drawn twice for clarity, initial states shaded, transient states dashed].

Since we deal with generally distributed session times when calculating the likelihoods of making future handoffs, we must track the number of completed handoffs. Thus, our transient chain consists of 2-tuple transient state definitions, where $(0, j)$ represents a session starting inside the j^{th} AGW, while the state (k, jr) is assigned to a session entering the j^{th} AGW through a border $r \in \{a, b, c, d\}$ after completing k handoffs. Similar to the transient Markov chain in Section 3.8.1, we also define two absorbing states: Z representing a session leaving the domain to the roaming partners and T representing the session termination. Fig. 3.17 shows a short summary of the model, where the initial states are shaded and the transitions to the absorbing states are dashed.

For a session starting in AGW j (i.e., state $(0, j)$), the transition probabilities are given by the joint event comprised of the transition probabilities $\hat{p}_{jl}$ for the initial movement (summarized in matrix $\mathbf{Q_{MI}}$), and the probability that the session contains at least one more handoff given that it made no handoffs, γ_1^x. For example in Figure 3.17, a session starting in AGW 1 (i.e., state $(0,1)$), leaving through border c and handing over to border a of AGW 4 (state $(1,4a)$) is described by the transition probability $\gamma_1^x \hat{p}_{1c}$. With (3.85) and (3.87), all transient transitions from $(0, j)$ to $(1, ir)$ are written in matrix form as $\mathbf{D_0} = \gamma_1^x \mathbf{Q_{MI}}$, where $\mathbf{D_0}$ is a $N_{\text{AGW}} \times 4N_{\text{AGW}}$ matrix. Similarly, the initial transitions

from transient states $(0,j)$ to the absorbing state Z are described by $\mathbf{A_0} = \gamma_1^x \mathbf{A_{MI}}$. Otherwise, if the session has terminated, the chain goes to the absorbing state T with a probability of $1-\gamma_1^x$. Now consider the example, that a session enters AGW 5 through border d after the second handoff (i.e., state $(2,5d)$ in Figure 3.17). If the session leaves the AGW through border b, the transition probability to neighboring border d of AGW 6 is given by $\gamma_3^x p_{5db}$. Using (3.85) and (3.87), all transient transitions out of states (k,jr) to states $(k+1,il)$ can be written in matrix form as $\mathbf{D_k} = \gamma_{(k+1)}^x \mathbf{Q_M}$, where $\mathbf{D_k}$ is a $4N_{\text{AGW}} \times 4N_{\text{AGW}}$ matrix. The transitions from transient states to the absorbing state Z are similarly described by $\mathbf{A_k} = \gamma_{(k+1)}^x \mathbf{A_M}$. Ordering states lexographically as $(0,1),...,(0,n),(0,1a),...,(0,nd),(1,1a),...,(1,nd),...,(k,1a),...,(k,nd)...$, the transient Markov chain is given similar to (3.76) as,

$$\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{D_0} & 0 & 0 & 0 & \cdots \\ 0 & 0 & \mathbf{D_1} & 0 & 0 & \cdots \\ 0 & 0 & 0 & \mathbf{D_2} & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \quad , \quad \mathbf{A} = \begin{bmatrix} \mathbf{A_0} \\ \mathbf{A_1} \\ \mathbf{A_2} \\ \vdots \end{bmatrix} \tag{3.88}$$

where $\mathbf{Q}$ is a matrix of unlimited size since the number of handoffs k can go to infinity. The elements of (3.88) are given as,

$$\begin{aligned} &\mathbf{D_0} = \gamma_1^x \mathbf{Q_{MI}} && , \mathbf{A_0} = \gamma_1^x \mathbf{A_{MI}} \\ &\mathbf{D_k} = \gamma_{(k+1)}^x \mathbf{Q_M} && , \mathbf{A_k} = \gamma_{(k+1)}^x \mathbf{A_M} \text{ , } k \geq 1 \end{aligned} \tag{3.89}$$

Let us denote the initial state probabilities for new on-net sessions as $\mathbf{P_I^{\Omega}}$ and emerging off-net sessions as $\mathbf{P_I^{\Phi}}$. The initial probabilities are then given as,

$$\mathbf{P_I^{\Omega}} = \left[\mathbf{F^{\Omega}}, 0, 0, ...\right] \quad , \mathbf{F^{\Omega}} = \left[\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{N_{\text{AGW}}}\right] \tag{3.90}$$

$$\mathbf{P_I^{\Phi}} = \left[\mathbf{F^{\Phi}}, 0, 0, ...\right] \quad , \mathbf{F^{\Phi}} = \left[\eta_{1a}, \eta_{1b}, \eta_{1c}, \eta_{1d}, \ldots, \eta_{N_{\text{AGW}}a}, \eta_{N_{\text{AGW}}b}, \eta_{N_{\text{AGW}}c}, \eta_{N_{\text{AGW}}d}\right]$$

where ε_j represents the probability of starting a session from AGW j for on-net traffic, and η_{jr} represents entering the network from edge r of AGW j for off-net traffic. It should be noted that $\mathbf{F^{\Omega}}$ is a $1 \times N_{\text{AGW}}$ row vector since on-net sessions initiate from within an AGW region while $\mathbf{F^{\Phi}}$ is a $1 \times 4N_{\text{AGW}}$ row vector since off-net sessions enter an AGW region from its borders.

Defining the fundamental matrix $\mathbf{M_z^{(x)}}$ as $\mathbf{M_z^{(x)}} = [\mathbf{e} - \mathbf{Q^{(x)}}]^{-1}$ and using the results from the transient Markov chain theory in (B.2), then similar to (3.83) the mean number of handoffs before leaving the network, $E\{K_x\}$ can be written as,

$$E\{K_x\} = \mathbf{P_I^{(x)}} \mathbf{M_z^{(x)}} \mathbf{o} - 1 \tag{3.91}$$

where $\mathbf{o}$ is an all ones column vector of the proper size. Using (B.3) and similar to (3.84), the roaming probability β_x is given as,

$$\beta_x = \mathbf{P_I^{(x)}} \mathbf{M_z^{(x)}} \mathbf{A} \tag{3.92}$$

Although (3.91)-(3.92) can be solved by matrix truncation as discussed in Section 3.8.1, this may result in large matrices and undesirable numerical errors. Next, we show a closed form result for the mean number of handoffs inside a network and relate our result to known work from [36] which does not consider the possibility of network departures and hence does not need to consider the mobility patterns between AGWs.

3.8.2.1 Vector Form Solution for the Mean Number of Handoffs

Using the residue theorem (see [36]), it can be shown that the mean number of handoffs inside the network, $E\{K_x\}$, can be written in closed form as,

$$E\{K_x\} = -\sum_{s_P \in \Xi_{\hat{S}-}} \operatorname*{Res}_{\hat{s}=s_p} \frac{f^*_{R_1}(\hat{s})\,\mathbf{F}^{(\mathbf{x})}\mathbf{Q}^{(\mathbf{x})}_{\mathbf{MI}} f^*_S(-\hat{s})}{\hat{s}} \mathbf{M}^{(\mathbf{x})}_{\mathbf{R}}(\hat{s})\,\mathbf{o} \tag{3.93}$$

where we have defined the matrix $\mathbf{M}^{(\mathbf{x})}_{\mathbf{R}}$ as,

$$\mathbf{M}^{(\mathbf{x})}_{\mathbf{R}}(\hat{s}) = \left(\mathbf{e} - f^*_R(\hat{s})\,\mathbf{Q}^{(\mathbf{x})}_{\mathbf{M}}\right)^{-1}$$

and that R_1 denotes the AGW residence time in the first AGW serving the session. Thus, $f^*_{R_1}(\hat{s}) = f_{\bar{R}}(\hat{s})$ for on-net sessions and $f^*_{R_1}(\hat{s}) = f_R(\hat{s})$ for off-net sessions.

It can also be shown that the roaming probability β_x, can be written in closed form as,

$$\beta_x = G^{(x)}(1)\,\mathbf{F}^{(\mathbf{x})}\mathbf{A}_{\mathbf{MI}} - \sum_{s_P \in \Xi_{S-}} \operatorname*{Res}_{\hat{s}=s_p} \frac{f^*_{R_1}(\hat{s})\,\mathbf{F}^{(\mathbf{x})}\mathbf{Q}^{(\mathbf{x})}_{\mathbf{MI}} f^*_S(-\hat{s})}{\hat{s}} \mathbf{M}^{(\mathbf{x})}_{\mathbf{R}}(\hat{s})\, f^*_R(\hat{s})\,\mathbf{A}_{\mathbf{M}} \tag{3.94}$$

Proof. see Appendix A.1.4 for the detailed proofs of (3.93)-(3.94). An example is also in Appendix A.1.4. □

To get an insight into (3.93), let us compare it the known result for the mean number of handoffs within a single network where the session never leaves its borders. It can be shown that the mean number of handoffs during the session is given as [36, 107, 108],

$$E\{K\} = -\sum_{s_P \in \Xi_{S-}} \operatorname*{Res}_{\hat{s}=s_p} \frac{f^*_{R_1}(\hat{s})}{(1 - f^*_R(\hat{s}))} \frac{f^*_S(-\hat{s})}{\hat{s}} \tag{3.95}$$

where the first residence time R_1 is usually assumed to be the residual of the residence time R. Comparing the mean number of handoffs in (3.95) and our result in (3.93), we observe that the scalar residence time terms $f^*_{R_1}(\hat{s})$ and $f^*_R(\hat{s})$ reflecting the speed of the

users and the AGW size are now multiplied by a spatial component which corresponds to the mobility model. $f^*_{R_1}(s)$ is multiplied by the initial movement matrix and the initial probabilities (i.e., $\mathbf{F^{(x)}Q^{(x)}_{MI}}$) and $f^*_R(\hat{s})$ is multiplied by the movement matrix $\mathbf{Q_M}$. Such elegant result allows us to say that the consideration of spatial aspects due to mobility patterns and AGW arrangements simply transforms the known solution for the mean number of handoffs from the scalar form in (3.95) to the vector representation in (3.93).

3.8.2.2 Exemplary Case Study

Let us now demonstrate the applicability of our model to obtain the mean number of handoffs for a session as a function of the mobility ratio, defined as $\rho = \frac{E_S}{E_R}$ [17] and the user mobility pattern. The mobility behavior is considered in two different topologies, linear and rectangular. For the linear topology, defined as $M_g = 1 \times N_g = 9$ we consider two different mobility patterns, referred to as "Random Route" and "Directed Route". The mobility pattern "Random Route" allows to change the direction at each AGW. Thus the matrix $\mathbf{Q_M}$ has nonzero entries $p_{1bb} = p_{9dd} = 0.5, p_{jbb} = p_{jbd} = p_{jdb} = p_{jdd} = 0.5, j = 2,..(n-1)$ and $\mathbf{A_M}$ is given by $[0, 0.5, 0, 0, ...0.5]^T$. In the mobility pattern "Directed Route", on the other hand, the user cannot change the initial direction, thus $\mathbf{Q_M}$ has elements $p_{jbd} = p_{jdb} = 1, j = 2,..(n-1)$ and $\mathbf{A_M}$ is given by $[0, 1, 0, 0, ...1]^T$. For the square topology, defined as $M_g = 3 \times N_g = 3$ based on the layout shown in Fig. 3.16, we consider the so-called "Random 3×3" and "Hotspot 3×3" mobility pattern. In "Random 3×3" mobility, all transition probabilities are set to 0.25. The initial probabilities $\mathbf{F^{(\Omega)}}$ for a new session is uniformly distributed. The mobility model "Hotspot 3×3" is defined by a combined random and directed movement pattern, as shown in the top part of Fig. 3.18, where a handoff session can choose a random direction only in AGWs 6-7. A new session starts with probability of 0.8 in AGW 5 and with equal likelihoods in the remaining gateways. For simplicity, we only consider on-net sessions.

As expected, the mean number of handoffs, MNH = $E[K_\Omega] + \beta_\Omega$, is always smaller than the mean number of handoffs for the whole session, i.e. for a network of unlimited size. The latter increases with increasing the mobility ratio as was also shown in [107, 108]. First, we observe that the "Random Route" mobility pattern results in a much larger number of handoffs than the 'Directed Route'. Compared to the directed movement, the random movement incurs several direction changes inside the network and since the user can only leave the network at AGW 1 or 9, a higher mean number of handoffs is observed due to the low roaming probability. However, this behavior highly depends on the topology and the mobility patterns. This is evident by comparing the 'Random 3×3' with the "Hotspot 3×3" mobility patterns. In this case, the random mobility behavior results in a much higher probability of roaming, β and thus the more directed Hotspot pattern achieves a higher mean number of handoffs. To sum up, Fig. 3.18 clearly shows that only by joint consideration of the network topology and the mobility pattern can the mean number of handoffs for a session be accurately estimated.

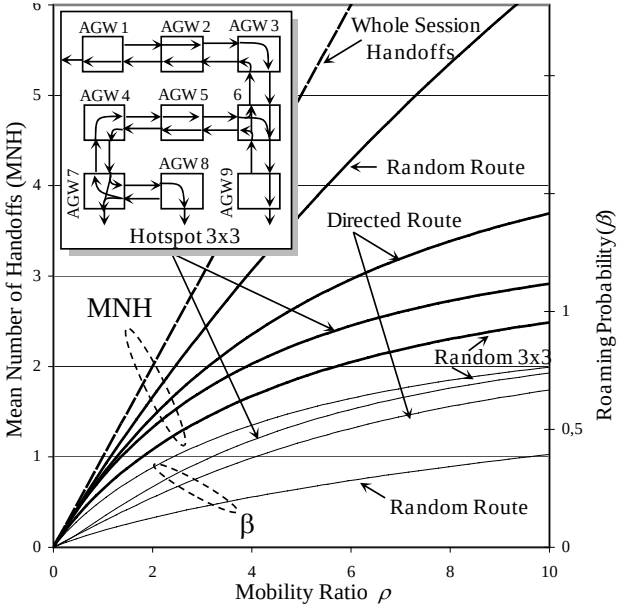


Figure 3.18: Mean number of handoffs and roaming probability β vs. mobility ratio for different mobility patterns (adapted from [125]) [session: Erlang, E_S=40 min, c_S = 0.5, residence time: Gamma, c_R = 2].

3.8.3 The Access Gateway Residence Time

In this section, we use the "cellular" residence time distribution (e.g., from measurements) to derive the AGW residence time statistics depending on the mobility model and network size. In our analysis, we assume that the cellular residence times, denoted as R_c, are independent and identically distributed (i.i.d). R_c is generally distributed with an existing Laplace transform and mean of E_{R_c}. We start by considering two exemplary cases based on directed (fluid flow) and random mobility patterns between cells. To keep the discussion simple, we assume linearly arranged AGWs, with $M_c \times N_c$ rectangularly arranged cells per gateway (see Fig. 3.19), and hence mobiles are only able to leave regions from the eastern and western borders. Next section, we outline how the pixel mobility modeling approach, which we used in the previous section, can be used to relax this assumption.

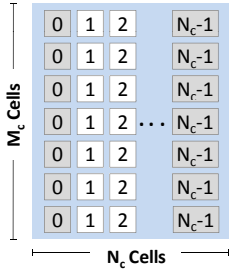


Figure 3.19: Cells within an AGW region (adapted from [127]).

3.8.3.1 The Case of Directed Mobility (Fluid Flow)

In this model, users move in a single direction, either east or west, with equal likelihoods and throughout their whole session duration. As such users move in the shortest possible path towards the boundaries of the AGWs. For the directed user movement the gateway residence time is composed of a sequence of cellular residence times R_c, which are assumed to be i.i.d. The resulting AGW residence time depends on whether we have an on-net session or an off-net session. When the session hands off from a neighboring gateway (i.e., off-net sessions), the mobile node may cross all N_C cells until it leaves the gateway, which yields the residence time R_g. On the other hand, when the session starts inside the gateway (i.e., on-net session), then the gateway residence times depend on the starting cell. The average overall possibilities of starting cells give us the residence time R_{g1}. Let us now proceed to obtain the residence time for on-net and off-net sessions.

Residence time for sessions starting inside the gateway (on-net sessions)
Let us first consider the case that the session starts in a cell that belongs to column $j-1, j \geq 1$ of a chosen AGW. Then if we first assume that the user moves to west he leaves the AGW after the j'th cellular handoff. For this case, the sequence of cellular residence time is given by the residual of the cellular residence time $\tilde{R}_c$ incurred in the first cell and $j-1$ subsequent cellular residence times. After j handoffs, the gateway residence time in the first AGW is then,

$$R_{g_1}(j) = \tilde{R}_c + \sum_{k=2}^{j} R_c \quad , \quad f^*_{R_{g_1}}(\hat{s}, j) = f^*_{\tilde{R}_c}(\hat{s}) \left(f^*_{R_c}(\hat{s})\right)^{j-1} \tag{3.96}$$

Assuming uniformly distributed session arrivals per cell and taking into account the symmetry for going west or east, the Laplace transform of R_{g_1} is

$$f^*_{R_{g_1}}(\hat{s}) = \frac{1}{N_c}\sum_{j=1}^{N_c} f^*_{\tilde{R}_c}(\hat{s})\left(f^*_{R_c}(\hat{s})\right)^{j-1} = \frac{1-f^*_{R_c}(\hat{s})}{\hat{s}N_cE\{R_c\}}\frac{1-\left(f^*_{R_c}(\hat{s})\right)^{N_c}}{1-f^*_{R_c}(\hat{s})} = \frac{1-\left(f^*_{R_c}(\hat{s})\right)^{N_c}}{\hat{s}N_cE\{R_c\}} \tag{3.97}$$

The AGW Residence Time for Handoff (Off-net) Sessions
The gateway residence time R_g for a session that handoffs from a neighbor gateway can only start in cell 0 and leave in cell N_C-1 or vice-versa. Due to the symmetry it follows that the AGW residence time and the Laplace transform of its PDF, $f^*_{R_g}(\hat{s})$, are given as,

$$R_g = \sum_{k=1}^{N_c} R_c \quad , \quad f^*_{R_g}(\hat{s}) = \left(f^*_{R_c}(\hat{s})\right)^{N_c} \tag{3.98}$$

Interestingly, we can observe here that (3.97) is equivalent to the residual of the AGW residence time in (3.96) as $R_{g_1} = \tilde{R}_g$.

3.8.3.2 Random Mobility Between Cells Within an AGW Region

In this section we assume that users move randomly between cells. To characterize the users' movements between cells, we use transient Markov chains. The transient states are used to model the cells inside the access gateway while the absorbing states represent departures from an AGW coverage area. For details on transient Markov chains, the reader is referred to Section B.1 and to [120].

In our context, as shown in Fig. 3.19, we view the access gateway as a collection of columnar groups of cells. After each cellular handover, the user may move to another cell within the same column i, leave the current column i and go east to column $i+1$, or go west to column $i-1$. Note that for simplicity of explanation we only consider mobility east-west between AGWs; the north-south mobility can be incorporated similar to [116] or more generally using the pixel movement approach as we will show later in this section. From the geometry, the probabilities of going east and west are equal and are $\alpha = 0.25$, while the probability of staying in the same column is $\zeta = 0.5$. We will model the user movement inside the AGW using a transient Markov chain. The access gateway area consists of N_c zones which represent the transient states. Only through two departure areas (i.e., shaded zones 0 and N_c-1) the user can leave the AGW. This is modeled by two absorbing states, G_E and G_W, representing departure to the east or west. The transition probabilities between the transient states are characterized by the

$N_c \times N_c$ matrix $\mathbf{Q_g}$ using the zone departure and stay probabilities α and ζ as,

$$\mathbf{Q_g} = \begin{bmatrix} \zeta & \alpha & \dots & & & \\ \alpha & \zeta & \alpha & 0 & \dots & \\ & & \vdots & & & \\ & & & & \alpha & \zeta \end{bmatrix} \tag{3.99}$$

The transitions to the absorbing states (i.e., departure east or west) are given by the $N_c \times 1$ column vectors $\mathbf{A_{gE}}$ and $\mathbf{A_{gW}}$

$$\begin{aligned} \mathbf{A_{gW}} &= [\ \alpha \quad 0 \quad \dots \quad 0\ ]^T \\ \mathbf{A_{gE}} &= [\ 0 \quad 0 \quad \dots \quad \alpha\ ]^T \end{aligned} \tag{3.100}$$

Residence Time for Sessions Starting Inside the Gateway (On-net Sessions)
Following our previous notation, when the session starts inside the AGW the residence time is R_{g1}. The new sessions can start with equal probability within any cell of the AGW. The initial state probabilities for the transient Markov chain are given as

$$\mathbf{f_{gI}} = [1 \dots 1]/N_c \tag{3.101}$$

Now the number of cellular handoffs until departure can be seen as equal to the number of transitions between transient states until absorbtion. Thus with (B.1), we can define the joint probability that the departure occurs with the j'th cellular handoffs to the east (or west) for new sessions as,

$$\Pr\{N_{G(I)} = j\} = \Pr\{N_{G(I,y)} = j\} = \mathbf{f_{gI}}\mathbf{Q_g}^{j-1}\mathbf{A_{gy}}, \tag{3.102}$$

where (I,y) specifies the starting state I and the departure side $y \in \{E,W\}$. Due to the symmetry we get the same distribution for both departure sides. Because $\mathbf{A_{gy}}$ contains only one absorbing state, (3.102) defines a discrete phase-type distribution, where j gives the number of time steps until absorption. Due to symmetry the probability of being absorbed into state G_W or G_E is given by $\beta_I = \beta_{IW} = \beta_{IE} = 0.5$. However, if we did not have uniformly distributed sessions as in (3.101), the absorption probabilities β_I would be calculated using (B.3).

Finally to derive the residence time we need the probability that the departure occurs with the j'th cellular handoff conditioned on the departure to the east or west, which simply follows from (3.102) as,

$$\Pr\{N_G = j | I\} = \Pr\{N_{G(I)} = j\}/\beta_I \tag{3.103}$$

Each step in the transient Markov chain represents a cellular residence time, where the first step has a duration of the *residual* cellular residence time due to the session start

($\tilde{R}_c$), followed by $j-1$ steps with duration of the cellular residence time (R_c). Similar to (3.97), the Laplace transform of the residence time, $f^*_{R_{g1}}(\hat{s})$, is given as,

$$f^*_{R_{g1}}(\hat{s}) = \sum_{j=1}^{\infty} f^*_{\tilde{R}_c}(\hat{s}) \left(f^*_{R_c}(\hat{s})\right)^{j-1} P\{N_G = j|I\} \tag{3.104}$$

$$= f^*_{\tilde{R}_c}(\hat{s}) \mathbf{f_{gI}} \sum_{j=1}^{\infty} \left(f^*_{R_c}(\hat{s}) \mathbf{Q_g}\right)^{j-1} \mathbf{A_{gW}}/\beta_I = f^*_{\tilde{R}_c}(\hat{s}) \mathbf{f_{gI}} \mathbf{M_g}(\hat{\mathbf{s}}) \mathbf{A_{gW}}/\beta_I$$

where we have defined the matrix $\mathbf{M_g}(\hat{\mathbf{s}})$ as,

$$\mathbf{M_g}(\hat{\mathbf{s}}) = [\mathbf{e} - f^*_{R_c}(\hat{s}) \mathbf{Q_g}]^{-1} \tag{3.105}$$

The mean residence time can be derived from (3.104)-(3.105) and is given as[11],

$$E\{R_{g1}\} = \frac{d f^*_{R_{g1}}(\hat{s})}{d\hat{s}} \mid_{\hat{s}=0} = E\{\tilde{R}_c\} + E\{R_c\} \left(E\{N_{G(I)}\} - 1\right) \tag{3.106}$$

where $E\{N_{G(I)}\}$ is the mean number of visits before absorption and is given as,

$$E\{N_{G(I)}\} = \mathbf{f_{gI}} [\mathbf{e} - \mathbf{Q_g}]^{-1} \mathbf{o}^T$$

Notice that the first visit has to be excluded, because it is not associated with a transition, i.e. a cellular handoff. The probability to leave is 1, thus $1 + (E\{N_{G(I)} - 1\})$ is equal to the mean number of cellular handoffs until leaving the AGW. Thus the mean residence time is composed of the residual cellular residence time $E\{\tilde{R}_c\}$ for the first cell and $(E\{N_{G(I)} - 1\})$ cell residence times $E\{R_c\}$.

The Gateway Residence Time ["Short" and "Long"]
Let us now analyze the details the gateway residence times for a handoff session entering the gateway region at any cell in edge columns (0 or $N_c - 1$) of the gateway AGW_k, $k = 0, ..., N_g - 1$. These sessions are particularly important because a handoff session staring in cell 0 will either handoff to the west AGW_{k-1} and thus present very short residence times R_{ga} within the gateway AGW_k ("short residence"), or on another extreme, the session may cross the whole gateway area, by moving east and experience very long residence times R_{gb}. The corresponding initial state probabilities are given as,

$$\mathbf{f_{gW}} = [\ 1 \quad 0 \quad \ldots \quad 0 \quad 0\ ]$$
$$\mathbf{f_{gE}} = [\ 0 \quad 0 \quad \ldots \quad 0 \quad 1\ ]$$

[11] The term $\frac{M_g(\hat{s})}{d\hat{s}}$ is given as,

$$\frac{M_g(\hat{s})}{d\hat{s}} = -M_g(\hat{s}) \frac{d\left([\mathbf{e} - f^*_{R_c}(\hat{s}) \mathbf{Q_g}]\right)}{d\hat{s}} M_g(\hat{s}) = M_g(\hat{s}) \mathbf{Q_g} \frac{d f^*_{R_c}(\hat{s})}{d\hat{s}} M_g(\hat{s})$$

Let us also denote the joint probability that the departure occurs with the j'th cellular handoff to the east (or west) given their initial starting edge east (west) as,

$$\Pr\{N_{G(E,E)}=j\}=\Pr\{N_{G(W,W)}=j\}=\Pr\{N_{G(a)}=j\}$$
$$\Pr\{N_{G(E,W)}=j\}=\Pr\{N_{G(W,E)}=j\}=\Pr\{N_{G(b)}=j\}$$

which is due to symmetry. Thus, the distributions of the number of handoffs are,

$$\begin{aligned}\Pr\{N_{G(a)}=j\}&=\mathbf{f_{gW}}\mathbf{Q_g}^{j-1}\mathbf{A_{gW}}\\ \Pr\{N_{G(b)}=j\}&=\mathbf{f_{gW}}\mathbf{Q_g}^{j-1}\mathbf{A_{gE}}\end{aligned}\tag{3.107}$$

Finally, the departure probabilities for a session starting at an edge zone and leaving at the same edge towards the neighboring gateway, β_{ga}, and the opposite case, β_{gb}, are

$$\beta_{ga}=\mathbf{f_{gE}}\mathbf{M_g}\mathbf{A_{gE}}=\frac{N_c}{N_c+1}\,,\ \beta_{gb}=1-\beta_{ga}\tag{3.108}$$

where $\mathbf{M_g}=[\mathbf{I}-\mathbf{Q_g}]^{-\mathbf{1}}$. The result for β_{ga} follows from the the last element of $\mathbf{M_g}$ derived by simple backward substitution. Similar to (3.103), we have to condition the probability distributions (3.107) by the departure probabilities, i.e.,

$$\Pr\{N_G=j|x\}=\Pr\{N_{G(x)}=j\}/\beta_{gx}, x\in\{a,b\}\tag{3.109}$$

For j cellular handoffs the access gateway residence time is given by $R_g(j)=\sum_{k=1}^{j}R_c$. Using (3.105), we get the Laplace transform for "short" and "long" residence times as

$$\begin{aligned}f^*_{R_{ga}}(\hat{s})&=\sum_{j=1}^{\infty}\left(f^*_{R_c}(\hat{s})\right)^j P\{N_G=j|a\}=f^*_{R_c}(\hat{s})\mathbf{f_{gW}}\mathbf{M_g}(\hat{\mathbf{s}})\mathbf{A_{gW}}/\beta_{ga}\\ f^*_{R_{gb}}(\hat{s})&=f^*_{R_c}(\hat{s})\mathbf{f_{gW}}\mathbf{M_g}(\hat{\mathbf{s}})\mathbf{A_{gE}}/\beta_{gb}\end{aligned}\tag{3.110}$$

Notice that (3.110) suggests that we have different residence times for a given AGW given where enter from. it follows that the observed residence time per gateway is a mixture random variable of the short and long residence times as,

$$f^*_{R_g}(s)=\beta_{ga}f^*_{R_{ga}}(s)+(1-\beta_{ga})f^*_{R_{gb}}(s)\tag{3.111}$$

3.8.3.3 Exemplary Application of the Model (Number of Cells per AGW)

This case is of particular importance due to the trend to have flatter networks with smaller sized distributed equipment (a.k.a AGWs)[23]. We consider a network that consists of a fixed number of cells as 10×120 and study the AGW residence time as function of the number of AGWs for directed and random mobility. For comparison

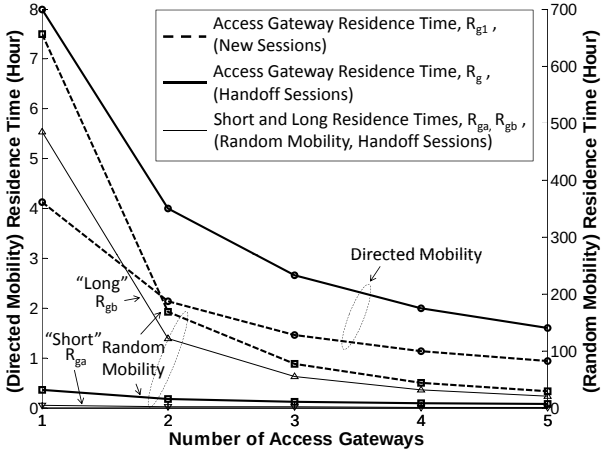


Figure 3.20: Mean gateway residence time vs. number of access gateways in an area (adapted from [127]) [$E[S] = 40$ min, $M_c = 10$, $E_R = 4$ min, $C_R = 2$, 5% offnet].

purposes, we study both movement patterns under at the same cellular residence time ($E[R_c]$= 4 min). We show results for the residence time incurred by new sessions R_{g_1} and for handoff sessions R_g derived in (3.96), (3.98), (3.104), and (3.111). We also show values for the two possible residence times for handoff sessions R_{g_a} and R_{g_b} derived in (3.110) as observed by random movers. As shown in Fig. 3.20, when the whole 10×120 cells are served by one gateway, the gateway residence times are very large. Directed movers always observe a relatively short residence time as they never change their movement direction, while random movers experience very long residence times as their movement behavior is relatively localized. We also show that for random users who make at least one handoff, they are most likely to incur R_{g_a} with a likelihood of $N_c/(N_c+1)$ after their first handoff and hence dominate the gateway residence time R_g.

3.8.4 Arbitrary AGW Residence Times and Mobility Patterns

In this section, we outline a hierarchical solution to estimate the mean number of handoffs under generic non-homogeneous AGW residence times and arbitrary mobility patterns by combining results from sections 3.8.2 and 3.8.3. As we discussed in Section 3.8.2, the pixel movement model can be used to describe arbitrary mobility patterns among arbitrary AGW areas with different sizes and residence times. The different residence times are due to the fact that the duration a user spends in an AGW area depends not only on the edge where she enters from but also on where she leaves from (e.g.,

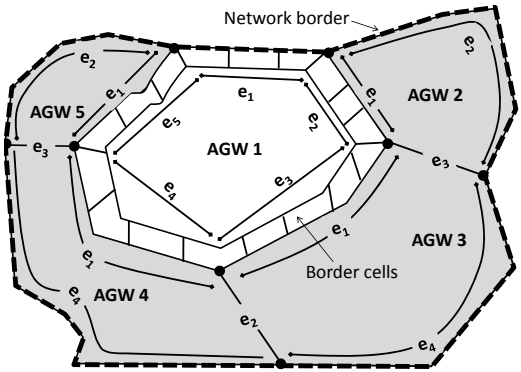


Figure 3.21: General cell layout within an AGW region.

see (3.110)). For analytical tractability, due to the non-homogeneous AGW residence times, the session duration is assumed to be *exponentially distributed*. A generalization for generic residence times can be performed as in [128]. Before we proceed, let us summarize the general steps involved in our analysis,

- Given the cellular arrangements, user concentrations $\mathbf{f_{gI}}$, and mobility patterns within each AGW region,
  1. Define N_e unique borders $e_1 \dots e_{N_e} \in \mathbb{E}$ for each AGW area. All cells lying on a border edge $e_i \in \mathbb{E}$ are associated with an absorbing state $\|\mathbf{A}\|_i$.
  2. For new sessions that initiate from within the AGW, calculate the probabilities of leaving from each border edge, e_i, denoted as $\|\hat{\mathbf{q}}\|_i$, as well as the corresponding Laplace transforms for the PDF of the AGW residence times given that the user left from border i as $\|\mathbf{f}_{\mathbf{R_{g1}}}(\hat{s})\|_i$. The matrices $\mathbf{f}_{\mathbf{R_{g1}}}$ and $\hat{\mathbf{q}}$ have $1 \times N_e$ elements.
  3. For handoff sessions, calculate the probabilities of leaving from each border edge, $e_j \in \mathbb{E}$, given that the user entered from border e_i where $e_i \in \mathbb{E}$, denoted as $\|\mathbf{q}\|_{(i,j)}$ as well as the corresponding Laplace transforms for the PDF of the AGW residence time when the user enters from border e_i and leaves from border e_j as $\|\mathbf{f}_{\mathbf{R_g}}(\hat{s})\|_{(i,j)}$. The matrices $\mathbf{f}_{\mathbf{R_g}}$ and $\hat{\mathbf{q}}$ have $N_e \times N_e$ elements.
- Having calculated the departure probability matrices and residence times for new and handoff sessions for each AGW in the network and given the initial session distributions in the network among AGWs, we directly apply the results in Section 3.8.2 to obtain the mean number of AGW handoffs in a network $E[K_x]$ as well as the probability of leaving the network $\beta^{(x)}$ where $x \in \{\Omega, \Phi\}$.

3.8.5 Models Limitations

The proposed models allow the estimation of the mean number of handoffs during a session as well as the network departure probabilities which are practically relevant to next generation all-IP network designs and are not investigated in the current cellular performance studies such as in [17, 37] and mobility literature for location management in cellular systems as in [117]. However, further investigation is still needed to reach a comprehensive solution in the following areas,

- *Correlated residence times:* As the network coverage areas might be relatively small, the consecutive residence times of cells and/or AGWs may be correlated. Such correlation may result in inaccuracies in estimating the mean number of handoffs. Up to our knowledge, correlated residence times were not considered in the literature so far.
- *The exponential session assumption:* The exponential assumption for the session duration in the generalized model in Section 3.8.3 may result in some inaccuracies as it allows only the consideration of first order statistics. Further investigation on the accuracy of the model and comparison with measured data is needed to estimate the error in the model's estimates.
- *The use of the residual session duration for off-net sessions:* In In some cases (e.g., due to higher tariffs for roaming), users may change their session duration statistics behavior and hence the session residual may not accurately capture reality. Discouragement factors should be incorporated to modify the session duration statistics according to cost when roaming.

3.9 Conclusions

In this chapter, we developed a generic analytical framework that allows the estimation of the AAA signaling load in mobile networks. Since the AAA signaling rate highly depends on the system configuration, we analyzed the load in three major AAA configurations, including fixed networks and mobile networks with a centralized and distributed AAA systems. The AAA model for fixed networks is derived for generic session distribution assumptions using stochastic and probabilistic methods. To accommodate mobility in the network, we extended the AAA model for fixed networks by utilizing concepts of holding and residence times from the cellular performance theory. While this approach is sufficient for centralized AAA systems primarily serving home users, it is inadequate to analyze scenarios with distributed AAA deployments and roaming. This is because users may move between areas served by AGWs reporting to different AAA systems or even between different network operators (i.e., roaming). We extended our analysis to track movements between regions in the network by combining our stochastic approach with a transient Markovian model. We use

transient Markov chains because we are not interested in mobiles' movements after they terminate their sessions. The session termination can be viewed as an absorbing state.

Our analysis clearly demonstrates that the AAA signaling rate is a nontrivial function of a spectrum of design parameters including AAA protocol settings such as the authorization lifetime and the accounting interim intervals, the authentication protocol, AAA system deployment (i.e., centralized or distributed), session statistics, mobility, and user densities in the network. Knowing the AAA signaling rate is critical as under provisioned systems can easily result in blocking users from access and losing valuable accounting information for the users' sessions. We showed that the signaling rate in all cases is a non-linear decreasing function of protocol settings such as the accounting interim interval and the authorization lifetime. We also showed that mobility in terms of the ratio of the mean session duration to the mean residence time linearly relates to the AAA signaling load in all deployments. We also showed that the mobility pattern affects both the number of handoffs as well as the likelihood of departing the network under consideration and hence results in non-linear effects on the observed signaling load. More detailed analysis on handoff modeling was provided in Section 3.8.

To sum up, the developed closed form results in this chapter can be useful for dimensioning AAA systems for various scenarios. As we will show in Chapter 4, they can also facilitate designing intelligent methods for optimizing AAA signaling in mobile environments. Although we have endeavored to cover many design variables and options, our models still require further enhancements to overcome further analytical and design challenges. These include the relaxation of the exponential session assumption in our analysis, the incorporation of signaling costs for each message based on measurements from currently available AAA packages (e.g., Free RADIUS [129] and Open Diameter [130] projects). In addition, a comparison based on measurements between RADIUS and Diameter protocols is needed to quantify the difference between these protocols as well as AAA designs running a dual stack of RADIUS and Diameter on the same platform. Furthermore, through sensitivity analysis is also required to determine the dependence of the models on estimated parameters including session, mobility, and users' distributions. Finally further investigation is needed to combine our models with forecast models for the service session and user growth in order to determine the frequency of system upgrades and the operators' operational expenses.

Chapter 4 Applications and Extensions

4.1 Introduction

In this chapter, we propose performance enhancements for the current AAA signaling mechanisms in order to mitigate QoS authorization delay and to enhance the accounting reliability in multi-service cellular networks. Specifically, we develop performance optimization techniques for QoS authorization for realtime services in IMS deployments (Section 4.3) as well as accounting reliability optimization for postpaid multi-service environments (Section 4.4). Afterwards, we propose two novel applications for AAA signaling in two important areas: (1) accounting for cellular backhaul services over wireless mesh networks (Section 4.5), and (2) inter operator layer 2 optical communications (Section 4.7).

4.2 Supporting Publications

1. S. Zaghloul, A. Jukan, "Optimal Accounting Policies for Reliability and Capacity of AAA Systems in Mobile Networks," to appear in the IEEE Transactions on Mobile Computing Journal, Jun 2009.
2. O. Tipmongkolsilp, S. Zaghloul and A. Jukan, "The Evolution of Cellular Backhaul Technologies: Current Issues and Future Trends," to appear in the IEEE Communications Surveys & Tutorials Journal, 1st Quarter, 2011.
3. S. Zaghloul, J. Aznar and A. Jukan, "Application Layer Signaling for Proactive Handoff Management in all-IP Wireless Networks," proceedings of the IEEE International Communications Conference (ICC'09), Germany, Jun 2009.
4. S. Zaghloul, W. Bziuk and A. Jukan, "A Scalable Billing Architecture for Future Wireless Mesh Backhauls," proceedings of the IEEE International Communications Conference (ICC'08), Beijing, China, pp. 2974-2978, Jun 2008.
5. S. Zaghloul, A. Jukan, "A Simple Signaling Mechanism for Seamless Interoperator Mobility in All-IP Networks," proceedings of the 5th Annual IEEE Consumer Communications & Networking Conference (CCNC'08), Las Vegas, USA, pp. 381-385, Jan 2008.

6. S. Zaghloul, A. Jukan, W. Alanqar, “Extending QoS from Radio Access to all-IP Core in 3G Networks - An Operator’s Perspective,” IEEE Communications Magazine, 45(9), pp. 124-132, Sep 2007.

7. S. Greco Polito, S. Zaghloul, M. Chamania and A. Jukan, “A Scalable AAA Signaling for Inter-carrier PCE Framework,” under submission.

4.3 Optimizing Handoff QoS Signaling Delay for Services

The recent evolution of wireless cellular systems towards all-IP architectures has stimulated unprecedented standardization and research efforts to support new service offerings, most notably within the IP Multimedia Subsystem (IMS) framework [22]. As we discussed in Chapter 2, services are authorized and their control of the offered QoS level is facilitated through interaction with the IMS policy function (i.e., the PCRF), which communicates the authorized QoS levels to the serving IP access gateway. When a mobility event is detected (e.g., mobile nodes cross the boundaries of gateway areas), IP layer signaling is triggered. Such signaling includes both Mobile IP and QoS authorization signaling with the service tier. While Mobile IP is needed to preserve IP connectivity, QoS signaling is needed to authorize the service delivery in the new access gateway region. The authorization time poses considerable challenges for mobile realtime services as it can vary considerably depending on several factors, such as service implementation, number of application servers hosting the service logic, and round trip delay between the cellular operator and the third party application providers [19, 131]. In general, authentication delay minimization was the subject of extensive research within academia and standardization bodies such as in the Media Independent Preauthentications (MPA) and the IEEE 802.21 Frameworks [8, 81, 132], in the context of Fast (proactive) Mobile IP handoffs as in [39, 80, 83, 133], and more recently in the context of Proxy Mobile IP where base stations act in the role of the Mobile IP client [82, 134]. However, little efforts focused on incorporating the delay relevant to the policy authorization in the service tier. In fact, only recently a discussion of such issues was made in the context of LTE using a preregistration procedure in [135].

In this section, we address the access gateway handoff delay due to QoS authorization by introducing a simple, application-layer proactive signaling mechanism that adapts to each service and its authorization delay requirements. In our method, the delay requirements of a service are passed from the service tier to the radio layer to assist handoff prediction, by leveraging the existing signaling systems for authentication, authorization, and accounting (AAA) systems. Let us now delve into the details of our mechanism in terms of signaling as well as service authorization delay estimation and handoff prediction.

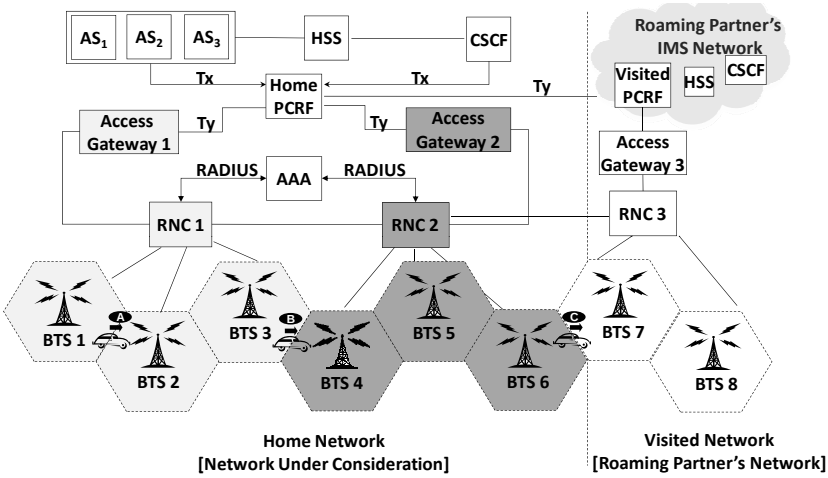


Figure 4.1: A simplified all-IP network architecture based on EVDO standards (adapted from [136]) [Acronyms used: BTS: Base Transceiver Station, RNC: Radio Network Controller, AAA: Authentication, Authorization, and Accounting, AGW: Access Gateway, CSCF: Call Session Control Function, PCRF: Policy Control and Charging Function, AS: Application Server].

4.3.1 Current Standards

Before we start describing our proactive mechanism, let us review the state of the art QoS signaling mechanism which we outlined in Chapter 2 and then proceed with details of our mechanism. We base our discussion on the 3GPP2 EVDO reference architecture shown in Fig. 4.1 based on [67]. Similar QoS signaling concepts also apply to the LTE standard (for details see [5, 137]). In our reference architecture, groups of cells are served by a radio network controller (RNC) and one or more RNCs are served by an AGW. Typically, a service is requested using Session Initiation Protocol (SIP) signaling between the mobile node and the service tier (i.e., IMS in our example). The serving call session control function (CSCF) is the first point of contact to the user and handles user registration and authentication by interacting with the home subscriber subsystem (HSS). The CSCF also routes SIP messages to application servers (AS) as well as to the called parties. In IMS, the home subscriber subsystem (HSS) contains the users' profiles including their service subscriptions. More detailed information on IMS can be found in [22].

When users move between RNCs belonging to the same access gateway (i.e, case A in Fig. 4.1), only radio layer handoff signaling is triggered and an optional authentication is carried out at the AAA server, typically through the RADIUS-based A12 interface

[40, 44]. However, when a mobile node moves between RNCs belonging to two different gateways (i.e, cases B and C in Fig. 4.1), the target gateway contacts the PCRF over the Diameter-based Ty interface [45, 138]. In this way, the target gateway can obtain the QoS profile for the handoff session in progress using the so-called service based bearer control (SBBC) signaling standard, see [67]. Depending on the service logic, the policy function may contact the CSCF and one or more application servers for QoS authorization, which can easily result in a handoff delay in the order of a second [131], which is obviously unacceptable for realtime services. This is because only when a response is received from the application servers can the PCRF respond to the target AGW over the Ty interface and authorize service in the target AGW's region. The handoff delay is further aggravated when users roam into other networks, i.e., visited networks, as illustrated in Case C in Fig.4.1. This is due to the fact that the visited network's PCRF needs to communicate with home network for authorization. In the following discussion the term handoff refers to the movement between two border base stations belonging to two different access gateway areas.

4.3.2 Enhanced Proactive QoS Signaling Mechanisms

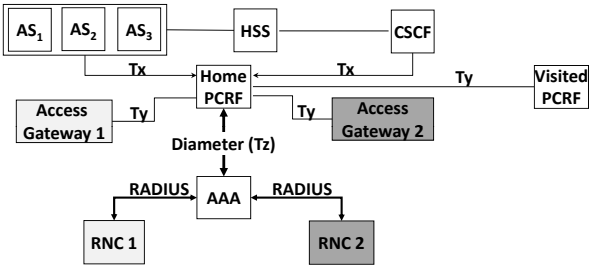


Figure 4.2: Proposed protocol interfaces.

To alleviate the QoS authorization delay during handoffs, we propose a proactive mechanism that pre-authorizes the session at the target gateway prior to handoff. For this proactive mechanism to be feasible, the service tier (i.e., IMS here) estimates the service authorization delays and conveys them to the radio layer, which in turn uses this information to predict handoffs. When a handoff is predicted within the radio tier, it informs the service tier about the predicted handoffs to start the QoS authorization signaling proactively towards the target AGW. In this way, our mechanism is able to lower the impact or potentially eliminate the handoff delay due to QoS authorization, while adapting to the service-specific authorization delays. To establish a communication path between the service and the radio layers, we benefit from the fact that both the PCRF and the AAA support the Diameter protocol and that the AAA has an existing interface

with the RNCs (i.e., the A12 interface) in the network. Hence, we only need to formally define a Diameter based interface between the AAA and the PCRF, which we call the Tz interface, akin to the SBBC parlance as shown in Fig. 4.2. We now present our protocol messages and the signaling flow. Afterwards, we explain delay estimation and handoff prediction implementation details.

4.3.2.1 The Protocol Messages

We propose that a new interface, referred to as Tz, is used between the policy function (PCRF) and the AAA system (see Fig.4.2). The Tz interface is easy to introduce as it uses Diameter protocol signaling already supported by both the AAA and the PCRF systems. It is implemented as a new authentication application and includes two primary messages: the Service Notification Request (SNRQ) and the Handoff Notification Request (HNR). The policy system uses the SNRQ message to inform the AAA about service authorization. On the other hand, the HNR message is sent by the AAA to inform the policy system of a probable handoff as soon as it receives a trigger from the radio layer. Note that the PCRF and the AAA system can serve as checkpoints; the services and their delays are inspected at the PCRF, while handoff indications are inspected at the AAA system. This minimizes the likelihood of instabilities due to mis-configurations.

We also use the already existing RADIUS based A12 interface [56, 57] between the AAA and the radio controllers to communicate handoff prediction triggers and service delay requirements between the radio and service tiers. Within the A12 interface, we define three new RADIUS messages, i.e, Service Authorization Latency Information (SALI), Handoff Imminent (HI), and Service Authorization for imminent Handoff (SAH). The implementation of our messages is based on RADIUS vendor specific attributes (VSAs) [40] carried in the authentication (access-request) messages. Since the SALI and SAH messages are server initiated, they are implemented similar to [139]. SALI messages are used to inform the border radio controllers within the gateway area about the delay requirements of the service authorization. SALI messages are only sent in case of a considerable change from the last delay measurement for a given service, or when the service is requested for the first time within an AGW region.

The information the SALI message carries is used by the handoff prediction algorithm in the RNC to send the HI messages to the AAA server indicating an imminent handoff. The SAH message is used to proactively authorize radio sessions at the target RNC prior to handoffs to eliminate the current A12 authentication delay. Since users may possibly be moving for long periods in the border areas between AGW regions, a large number of HI messages can be incurred. To address this issue, the HI message authorizes the session at the target gateway for a predefined *authorization interval*. The authorization interval should be chosen based on the tradeoff between low signaling load and the reservation of the radio layer and memory resources at the RNC.

4.3.2.2 The Proactive Signaling Flow

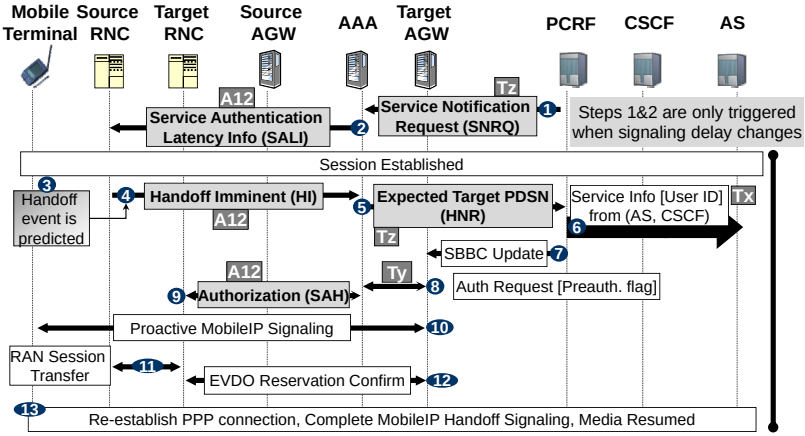


Figure 4.3: Proactive signaling flow (adapted from [136]).

Fig. 4.3 shows the corresponding signaling flow. In step 1, the policy system provides an estimate of the authorization delay to the AAA server (SNRQ), which is followed by a SALI notification to the RNCs (step 2). Notice that these messages are not sent on a per session basis but rather when an appreciable change in the service authorization delay is observed. When the RNC determines the likelihood of an imminent gateway handoff (step 3), it sends a HI message including the estimated handoff time to the AAA server (step 4). The AAA system may prioritize the processing of HI messages based on service priorities and inform the policy system of the imminent handoff using the HNR message which carries the expected target AGW information (step 5). In step 6, the policy system requests QoS authorization information for the session from the application servers and the CSCF. It also checks its local policies for the target AGW. If successful, the policy system informs the target AGW about the imminent handoff using SBBC signaling [67].

The target AGW then requests pre-authorization from the AAA by setting a vendor specific pre-authentication attribute (step 8). The AAA authorizes the request and optionally preauthorizes the request at the target RNC by sending the SAH message (step 9). In steps (10-13), the standard (proactive) Mobile IP [39, 80, 83, 133], radio flow reservation, and point-to-point (PPP) connection establishment are performed. Once Mobile IP handoff completes, it is unnecessary to authenticate at the service tier and hence the media session resumes with minimal delay.

4.3.3 Delay Estimation and Handoff Prediction

4.3.3.1 Authorization Delay Estimation

The policy system (PCRF) maintains an average estimate of the authorization delay per service (W_s). Since the Tx interfaces between PCRF and CSCF as well as between PCRF and AS are based on Diameter protocol, estimates of the signaling delay can be obtained from the time an authorization request is sent until an answer is received, or from the Diameter watchdog messages when the interfaces are idle. If the n$^{\text{th}}$ authorization delay W_s^n differs from the last estimate by a given margin, δ, an update is sent to the underlying radio network through the AAA framework (see Mechanism 1).

Mechanism 1 IMS Authorization Latency Update Mechanism

Input : Set of services, Set of border RNCs
Output : The authorization latency for each service
foreach *update step n* **do**
  foreach *AGW i* **do**
    foreach *Service s* **do**
      Measure D_s^n as the authorization delay to application servers and CSCFs.
      Measure $D^{(n,i)}$ as the round trip delay from each AGW i to the PCRF.
      Calculate $D_s^{(n,i)} = D_s^n + D^{(n,i)}$
      Calculate the moving average $W_s^{(n,i)}$ using $D_s^{(n,i)}$
      if $\mid W_s^{(n,i)} - W_s^{(n-1,i)} \mid > \delta$ **then**
        Send a SNRQ to all AAA servers
        All AAA servers update all *border* RNCs using the SALI message
        All border RNCs use the new $W_s^{(n,i)}$ for prediction
      end
    end
  end
end

The margin δ is a critical parameter to the stability of the system as it determines the frequency of the SALI messages and hence the stability of the handoff prediction. Since δ is highly dependent on the AS loading, it is important to select a margin such that the moving average, W_s^n, is stable and is barely affected by the server load fluctuations. To illustrate this effect, let us assume that all N_s ASes hosting the service logic incur similar loading and that the policy function forks its authorization requests to the application servers. Since each AS responds after a random delay d, the authorization latency D_s^n is determined by the latest responding server (i.e., $D_s^n = \max\{d_1, ..., d_{N_s}\}$). Assuming quasi-stationarity and by central limit theorem, W_s^n is normally distributed with mean $E[D_s^n]$ and variance of $\text{Var}[D_s^n]$/(Window Size). Assuming M/M/1 application servers, typical values are obtained for δ as in Table 4.1. We observe that the required margin grows approximately linearly with the AS load until loads of 80% and exponentially afterwards. Depending on the service time of the AS, one can select a suitable value for δ (e.g., if the AS serves 50 req/s, then $\delta = 3.84(20) = 76.8$ms at 90% load).

Table 4.1: The margin δ normalized to the mean service duration of the application server as function of its loading

Load	**10%**	**25%**	**50%**	**75%**	**80%**	**85%**	**90%**	**95%**
δ **(1 AS)**	0.37	0.44	0.66	1.32	1.65	2.19	3.29	6.58
δ **(2 AS)**	0.41	0.49	0.74	1.47	1.84	2.45	3.68	7.36
δ **(3 AS)**	0.43	0.51	0.77	1.54	1.92	2.56	3.84	7.68

Mechanism 2 Handoff Prediction Mechanism

Input : The *TrackedSet* $\mathbb{B}_{\mathbb{T}}$, the authorization delay $W_s^{(n,i)}$
Output : Handoff Imminent (HI) message
/* This logic runs at each time step, n (e.g., 100ms) */
foreach $BTS_j \in \mathbb{B}_{\mathbb{T}} \cap (\mathbb{B}_{\mathbb{A}} \cup \mathbb{B}_{\mathbb{C}})$ **do**
  Compute the mean signal to interference ratio $\left(\frac{E_b}{N_o+I}\right)_j$
  $\Delta_j = (E_b/(N_o+I))_m - (E_b/(N_o+I))_j$
  Estimate the rate $R_j = d\Delta_j/dt$ using the last M samples
  if $R_j < 0$ *AND* $E_b/(N_o+I)_m \leq$ *Threshold* **then**
    $\Delta_0 = \Delta_H + \Delta_j$, $T_j = |R_j|^{-1}\Delta_0$
  end
end
$T_k = \min\{T_j\}$
if *NOT IsTriggered AND* $T_k \leq W_s^{(n,i)}$ **then**
  IsTriggered = true
  Start Timer = *Authorization Interval*
  Send HI message including T_k to the AAA system
end

4.3.3.2 Handoff Prediction

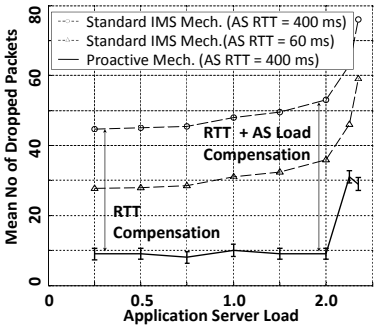
The authorization delay estimate $W_s^{(n,i)}$ for AGW_i is used by the border radio network controllers to predict handoff events and therefore attempt to trigger the proactive authorization process $W_s^{(n,i)}$ seconds prior to the estimated handoff instant. In this section, we use a simple linear prediction to estimate the handoff instant as shown in Mechanism 2. By monitoring the candidate $\mathbb{B}_{\mathbb{C}}$ and the active sets $\mathbb{B}_{\mathbb{A}}$ of base stations for each mobile node, RNCs are able to predict handoff moments; a candidate set includes base stations with received powers below a certain threshold, and once this threshold is exceeded they are added to the active set. Let us define the *TrackedSet*, $\mathbb{B}_{\mathbb{T}}$, as the set of bordering base stations in a border RNC within an AGW coverage area. This is needed because not all base stations in RNC regions are at the edge. The handoff moment can then be estimated by monitoring the power decay rates R from all base stations belonging to the *TrackedSet* that appear either in the candidate or active sets (i.e., $BTS_j \in \mathbb{B}_{\mathbb{T}} \cap (\mathbb{B}_{\mathbb{A}} \cup \mathbb{B}_{\mathbb{C}})$) as shown in Mechanism 2. Then, the handoff moments can

be estimated by the products of the decay rates and the sum of the handoff hysteresis margin Δ_H and the difference between the signal to interference ratios Δ_j from the current and the target base stations (i.e., $(E_b/(N_o+I))_m$ and $(E_b/(N_o+I))_j$ respectively). Once estimates of the handoff moments are collected, the earliest predicted handoff event from base station k, denoted as T_k , is compared to the estimated authorization delay $W_s^{(n,i)}$ and if $T_k < W_s^{(n,i)}$, a handoff imminent (HI) message is sent to the AAA server reporting the possible source and target RNCs and the authorization lifetime is set to T_A. During the authorization lifetime, no HI messages are sent by the source RNC in order to prevent continuous triggering of HI messages. The complexity of the prediction mechanism grows linearly with the number of base stations and services.

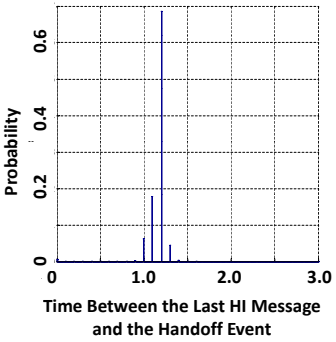
4.3.4 Mechanism Evaluation

In this subsection, we show results that demonstrate the correct operation of the mechanism while we leave performance related discussion to Chapter 5. To do so, we focus on the interplay of our signaling and the data plane traffic. We study an exemplary VoIP application and the number of dropped VoIP packets during gateway handoffs using our proactive signaling mechanism. To evaluate the handoff impairments due to the service tier QoS signaling, we assume *an ideal* proactive Mobile IP handoff delay of 140 ms and a typical 70 ms delay in the radio layer based on EVDO technology. We monitor the mean number of dropped VoIP packets during handoffs for various AS loads ranging from 40% to 95%. More details on the simulation setup is described in Chapter 5 and in [140].

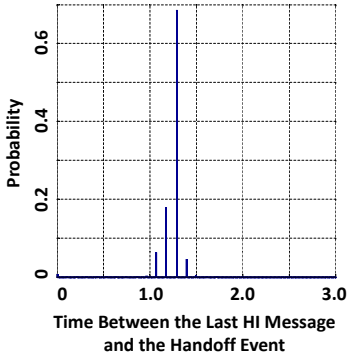
As shown in Fig.4.4(a), we see that our proactive mechanism is able to minimize the number of dropped packets even for relatively large round trip delays between the PCRF and the AS. We also observe that the number of dropped packets in the standard IMS mechanism is sensitive to the round trip time to the ASes as well as their load. When the AS load exceeds the 90% limit, the variance of the authorization delay at the AS increases rapidly and our prediction mechanism is no longer able to correctly adapt leading to a large number of packet drops. This effect is clear in Figs.4.4(b)-4.4(d) where we plot the probability mass function of the time between the last HI message and the handoff trigger. Notice that due to the shape of the histograms in Fig.4.4(b) (i.e., 40% loading) and Fig.4.4(c) (i.e., 90% loading) are similar while the delays start to "spill out" when the AS is increased to 95% due to the large variance in the authorization latency and hence resulting in improper prediction of the handoff time. We also observe that the largest component in the histograms shifts according to the AS load (i.e., 1.25 in Fig.4.4(b), 1.40 in Fig.4.4(c), and 1.75 in Fig.4.4(d)) and hence explains the flat shape of the number of packet drops in Fig.4.4(a).



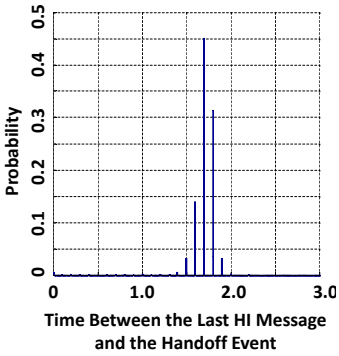
(a) Avg. dropped VoIP packets during handoff



(b) PDF of the time between the last HI message and the handoff event [AS Load = 40%]



(c) PDF of the time between the last HI message and the handoff event [AS Load = 90%]



(d) PDF of the time between the last HI message and the handoff event [AS Load = 95%]

Figure 4.4: The performance of a VoIP stream using our method compared to standard IMS schemes (adapted from [136]) [Authorization Interval = 150s, δ set at 90% AS loading].

4.3.5 Open Issues

Although our scheme pro-actively eliminates the signaling delay in several cases, further research is still needed in the following areas,

- *Integration with emerging standards*: Our work can be integrated with the emerging 802.21 media independent handover framework to mitigate the IMS QoS signaling delays within the information, command, and event services [132]. This can be achieved by integrating our work with the media independent pre-authentication (MPA) standardization efforts which currently focuses on pre-authentication due to Mobile IP signaling and does not incorporate the PCRF from the IMS layer yet. Our proactive QoS signaling work can also complement the proactive PMIP signaling proposed for vertical handoffs between different technologies which only consider the PCRF QoS signaling in principle during a pre-registration phase [135].

- *Supporting multi-bearer communications over multiple systems*: In emerging systems, it is possible that a rich session be served by multiple technologies such as UMTS serving voice and LTE carrying video streams. In such cases, the network selection [141] should be considered in order to properly pre-authorize for the proper QoS for the bearer in the target networks or gateway regions.

- *Investigation of mixed hierarchical and proactive policy methods*: Hierarchical designs for the policy framework were proposed in [9] to reduce the signaling load and to speed up QoS authorization within a domain. This is achieved by delegating the QoS authorization to local PCRFs within domains by a central PCRF. However, this raises latency issues when application servers require that QoS be negotiated during handoff moments or when users roam between networks. Further research is needed to combine proactive and hierarchical QoS signaling in policy systems by delegating QoS signaling ahead of the handoff moments, say at session initiation instants or within an operator domain. In all cases, our proactive QoS signaling solution is needed for inter-operator signaling or when application servers do not allow delegating QoS authorization.

- *Supporting policy interworking functions*: In some cases, such as when users move between LTE and WiMAX systems, QoS interworking signaling is needed to convey QoS parameters and charging rules to the target network [142]. This operation can take extra time and should be considered in proactive QoS signaling frameworks.

- *Investigation of integration with prepaid systems*: Integrating QoS proactive signaling for prepaid users [51] is non-trivial as charges may vary depending on the target gateway area or network. Complexities can arise due to the need to always confirm that the user has enough credits until the handoff actually takes place.

4.4 Accounting Optimization for Multi-Service Postpaid Systems

The success of next generation IP-based mobile systems in terms of the operator's revenue growth largely depends on the abilities to implement smart charging and accounting strategies for the supported Quality of Service (QoS). As we know from previous chapters, the accounting interim records are issued periodically during the service sessions as means of protection against server or network failures or even loss of accounting stop messages. Clearly, unreported usage can lead to a significant loss of revenue [7, 45]. For instance, for a typical size equipment [38], the failure of a NAS serving 24,000 active users from 800 base stations with average session duration of 10 mins and a charge of 10 cents a minute, results in a loss of 12,000 USD when the reporting interval equals 10 mins. A reduction of the potential loss by half by reducing the reporting intervals, would result in requirements to handle about 30% more signaling load; a further loss reduction to 1,000 USD would require the signaling server capacity to go up to 314%. Clearly, there is a tradeoff between the potential loss and the signaling load; the shorter the reporting interval the smaller the potential loss, but also the larger the signaling load and hence the required size of the AAA system [7]. As the current AAA standards [40, 41, 45] leave the determination of the reporting periods open to the operators, the question arises of how to minimize the potential losses while avoiding excessive server over-provisioning, especially as the number of mobile services is expected to grow and energy and data center sizes are becoming a concern [143].

Finding an optimal tradeoff between the potential loss and the signaling load is particularly complex in mobile and multi-service network systems, as the multi-service and mobile scenario results in a multi-commodity trade off due to the potential loss from each service, its session statistics which vary with mobility, and the corresponding signaling load from all services. The impact of mobility is non-trivial. For mobile services, the optimality for the reporting periods can only be achieved by adapting the reporting intervals to the expected service session arrival rates, service durations, and their costs. These expected values vary and often do not exhibit long term stationarity. For some mobile users, only a portion of the session is observed by the serving NAS. Depending on the users' concentration in the border areas of the cellular coverage area under consideration, the service sessions arrival rates and their effective service time within the NAS area may also largely fluctuate. Hence, even though operators can choose to determine the reporting intervals empirically and based on past observation, future services can be better served by a formal characterization of the accounting intervals which can optimally relate signaling load to the potential loss.

To address these issues, in this section we propose the first formal framework that quantifies the trade off between the potential loss and the signaling load in multi-service mobile networks. We also propose two optimization policies, which can adaptively and optimally trade off the potential loss and the AAA signaling load. In our framework, we utilize stochastic and renewal theoretic concepts to obtain simple estimates of the

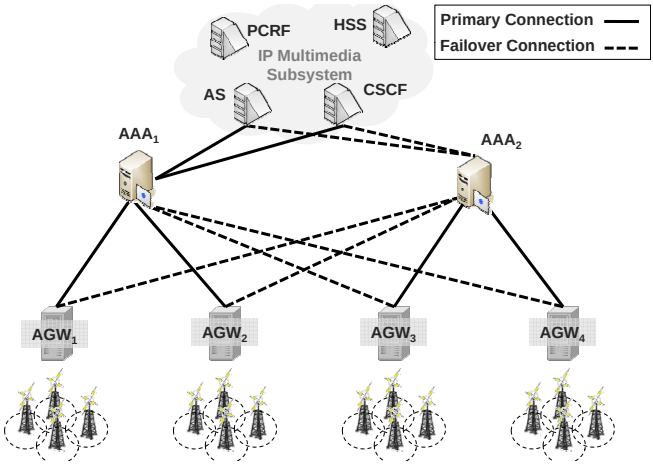


Figure 4.5: Simplified system architecture (adapted from [145])

signaling load and the potential loss to be used by the optimization policies. To account for statistical variability due to mobility, our method uses standard protocol attributes [43, 44] to categorize mobile service sessions into four distinct types relevant to their initiation and termination locations. The statistics of the four components are then used to estimate the load and loss by extending concepts of session holding time, based on the models we developed in Chapter 3. The proposed optimization mechanism embraces the current AAA IETF standards RADIUS and its successor Diameter and does not require any modifications to the AAA protocols nor to the network access servers' implementation and its implementation scope is limited to the AAA systems. As such the method is easy to implement and scalable with the number of services. Finally, our proposed work is different from other efforts in this area as their primary focus was dedicated to other aspects such as service metering configuration and management [35], enhancements to accounting schemes in high mobility networks [144], and challenges for fraud detection as in [49].

4.4.1 Overview of the Optimization Mechanism

Figure 4.5 shows a simplified all-IP wireless network architecture which consists of four access gateways serving four cellular regions[1]. The four AGWs connect to two AAA systems in a redundant pair configuration. IMS network elements are also shown

[1]Recall that an AGW is a generic term that refers to the first IP network element which interacts with the terminal and usually implements the network access server (NAS) functionality.

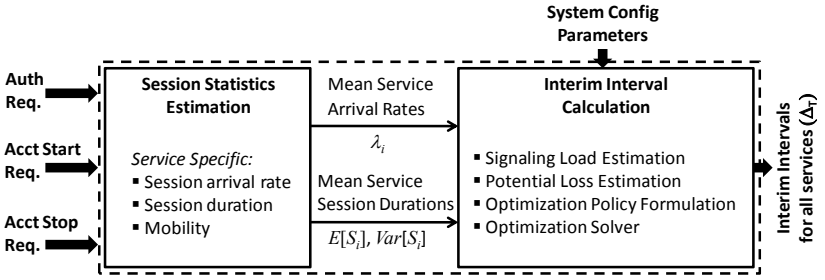


Figure 4.6: The mechanism's block diagram (adapted from [145]).

in the Figure 4.5 including the Call Session Control Function (CSCF) which acts as a soft switch, Application Servers which host the service logic, the PCRFs which provide the AGWs with QoS authorization and charging information for the requested services, and the Home Subscriber Subsystem (HSS) which hosts the users' profiles. Hence, service flows are identified by AGWs using charging rules supplied by the PCRF [22] and metered accordingly. Hence, the accounting process is referred to as flow based accounting as explained in Chapter 2. The accounting signaling process follows the standard procedure described in Section 3.5.

In this section, we use the term *service session* as a generic term which refers to duration of the chargeable service flows rather than the mere connectivity time at the IP level. It is noteworthy to state that currently RADIUS and Diameter only support time based interim reporting. Recent proposals [54, 146] have suggested the triggering for interim records based on consumed data volumes for data based services, (e.g., after 500KB of data are consumed by a terminal). When volume based interim reporting is possible, our method can be directly applied by using volume rather than time units as the distribution of the packet volumes that were transmitted in a service session can be mapped to a specific service session holding time distribution [86]. For the rest of this section, we will focus on time based metering and use the NAS as a general term to refer to the AGW, CSCFs, or ASes.

Figure 4.6 shows a high level diagram of the proposed optimization mechanism. Our scheme can be viewed as an AAA module which receives the authentication accounting start and accounting stop requests and use them to update the accounting interim intervals from all services that will be used by currently arriving and future service sessions. Our mechanism consists of two major blocks: one responsible for estimating service load and session duration statistics and another that uses such estimates to resolve the tradeoff between the load and the potential loss to produce optimal interim intervals for all services based on the current state of the system. We emphasize that our scheme is not an overload handling mechanism but rather targets resolving the tradeoff between the loss and the load and leaves the overload handling mechanism intact. Since ac-

cording to the RADIUS and Diameter standards [41, 45], it is generally not possible to change the interim intervals for the admitted sessions, the optimized interim settings only affect future sessions.

In a nutshell, the statistics estimation block tracks the current service session arrival rates, duration, and mobility statistics from all NASes. When a sufficient change in the service session arrival rate or duration statistics or a change in the system's parameters is detected, interim recalculation is invoked. In this regard, the estimates of the potential loss and the signaling load are updated based on the estimated statistics, which are then used along with configuration parameters by the optimization policies. The optimization policies are then solved by the optimization solver and the interim intervals are updated based on the latest state of the system. The typical triggers of a new statistic estimation can be tariff switching, NAS failover, NAS addition or removal, but triggers can also be periodic, for administrative reasons. The interim interval calculation also considers the configuration parameters. For our mechanism, each service's configuration includes the administrative range for the interim intervals [7] denoted in vector form as $\mathbf{\Delta_T^{min}}$ and $\mathbf{\Delta_T^{max}}$ and service costs. The configuration parameters also include the capacity of the AAA system P, and whether optimization is allowed. The last parameter is useful in cases where the interim interval for some services is required to be fixed such as in some roaming agreements or for administrative reasons. In the following discussion, we provide details on each functional block shown in Figure 4.6.

4.4.2 The Session Statistics Estimation Block

The major functions of the statistics collection block is to keep track of the services session arrival rate and duration statistics (e.g., mean and variance) including mobility effects, and then trigger an interim interval recalculation when the service statistics change by an amount greater than a preset threshold.

4.4.2.1 The Service-specific Session Statistics

In our system, each service is identified by unique properties such as NAS IP address, service type (e.g., VoIP, video, gaming, etc), cost, etc. We use moving average windows to maintain the most recent statistics for the arrival rate and the session duration of each service served by the AAA system. The moving windows are used to adapt to changes in service statistics during the day. The collected statistics for each service include the access request rate, the rejected authentications rates (e.g., mis-configured devices), and session durations. In practice, this is possible as many of the available AAA solutions today already implement traffic counting abilities and offer them for network operations and management systems [147, 148]. The mean session arrival rate is estimated by the inter-arrival time between accounting-start requests and the session duration is directly read from the `Session-Time` attribute in the accounting stop messages. To account for

mobility effects, other attributes are used as we described in Table 2.1. The estimated mean arrival and session durations for each service are used to trigger a recalculation of the interim interval when a change in the mean arrival rate or session durations exceeds a preset threshold (e.g., 5% since the last interim optimization). To ensure resilience against transient perturbations in service statistics, we also wait for a minimum grace period to pass since the last optimization operation.

4.4.2.2 Impact of Mobility on Session Statistics

When users move between NAS regions, the accounting sessions are closed on the source NAS (i.e., access gateway) and new accounting sessions are started at the target NAS. Consequently, this has an impact on the session statistics observed at the AAA system from a particular NAS. To capture this important aspect, we use the AGW holding time which denotes the duration a service session spends in a given NAS region before it terminates or moves to another NAS area as we elaborated in Section 3.5.3. In short, this definition leads to four basic service session holding time categories, as illustrated in Fig. 4.7, i.e.,

1. *Full Sessions,* $H^{(F)}$: Sessions that originate and terminate in the NAS area under consideration.
2. *Originating Sessions,* $H^{(O)}$: Sessions that originate in the NAS area under consideration and last long enough to handoff to other NAS serving areas.
3. *Terminating Sessions,* $H^{(T)}$: Sessions that originate in another NAS area and terminate in the NAS area under consideration.
4. *Transit Sessions,* $H^{(Tr)}$: Sessions that pass through the NAS area under consideration (i.e., start and terminate in other NAS areas).

Notice that for NAS 1 in Fig.4.7, mixed mobility cases such as case 5 can be decomposed into cases 2 and 3 and hence do not need to be addressed separately. Thus, our characterization is sufficiently general to handle both fixed and mobile systems. For instance, depending on the size of the NAS area and its surrounding NASes, different behaviors can be observed. For instance, for networks with large NAS areas, handoffs are unlikely and hence $\lambda_i^{(F)}$ is high and hence the relative proportions of $H_i^{(F)}$ dominate. If the NAS under consideration was large and surrounded by small NAS areas then $\lambda_i^{(F)}$ and $\lambda_i^{(T)}$ will be large and hence $H_i^{(F)}$ and $H_i^{(T)}$ will dominate. Similar arguments can be made when significant user concentrations are located in its border cells of the NAS coverage area.

The protocol attributes necessary to obtain the four session holding times based on RADIUS and Diameter are shown in Table 4.2 based on our discussion in Chapter2. The

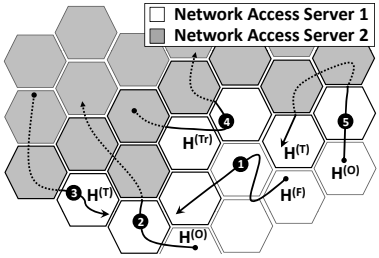


Figure 4.7: AGW holding times (solid lines) from the perspective of NAS 1 for various mobility patterns (adapted and modified from [145]) [session types: (1) Full, (2) Originating, (3) Terminating, (4) Transit, (5) Mixed (originating and terminating)]

attributes in the table are used in several wireless systems such as WiMAX and 1xEVDO systems [43, 44]. Recall that the `Beginning-of-Session` attribute is used to mark the first accounting period in a session and appears only in accounting start messages. The `Session-Continue` appears only in accounting stop messages and is used to indicate whether there are any subsequent accounting periods. The session holding times can be read directly from the accounting stop records from the standard `Acct-Session-Time` [41, 45] attribute which reports the service time for the session by a particular NAS element. Thus, for each service i, the output of the statistics estimation block is given as four components for the session arrival rates, $\lambda_i^{(x)}$, the four components for the session holding time, $H_i^{(x)}$ where $x \in \{F, O, Tr, T\}$, and service authentication success rates.

Table 4.2: Session types categorization using RADIUS/Diameter AVPs [145] [Acronyms, AVP: Attribute Value Pair, BOS: Beginning-Of-Session, SC: Session-Continue]

Session Type	BOS AVP [Acct. Start]	SC AVP [Acct. Stop]
Full, $H^{(F)}$	true	false
Originating, $H^{(O)}$	true	true
Terminating, $H^{(T)}$	false or N/A	true
Transit, $H^{(Tr)}$	false or N/A	false

4.4.3 The Load and Loss Estimation

In the interim interval calculation block, services are grouped into NAS sets, denoted as $\mathbb{N}$, which identify all service sessions coming from the same NAS node. This is needed for the loss estimate because failures usually impact one NAS and not all NASes simultaneously. The global service set, $\mathbb{A}$, used to estimate the signaling load, is the

union of all NAS sets and is given as $\mathbb{A} = \mathbb{N}_1 \cup \mathbb{N}_2 ... \cup \mathbb{N}_k$ where $\mathbb{N}_k$ is the k^{th} NAS set.

4.4.3.1 Estimating the AAA Signaling Load

In this part, we shortly review concepts we developed in Chapter 3 for the signaling load. Let us assume the generic case that the AAA signaling traffic consists of both authentication and accounting messages, otherwise the authentication terms are simply ignored. For clarity, we first review the estimation of the signaling load in the absence of mobility, as was shown in the fixed model derived in Section 3.4, and then show how to incorporate mobility effects using the concept of the holding times used in Section 3.5.3. Let us denote the mean AAA signaling rate as ξ. Let ξ_A, ξ_R, ξ_{Start}, ξ_{Int}, and ξ_{Stop} denote the mean authentication,reauthentications, accounting start, interim, and stop rates respectively. Let p_a denote the estimated AAA authentication success rate probability (i.e., the estimated proportion of the accepted access requests). Let us also assume that the service session arrival process is Poissonian. Then from Section 3.4, the resulting signaling rate is then the sum of all the rates from all services including authentications, accounting starts, interims, and stops and is given as,

$$\xi = \sum_{i \in \mathbb{A}} \left[\xi_{A,i} + \left(\xi_{R,i} + \xi_{Start,i} + \xi_{Int,i} + \xi_{Stop,i} \right) p_{a,i} \right] \tag{4.1}$$

In (4.1), we make the assumption that reauthentications are always successful for already authenticated users. Following a similar approach as in Section 3.4, the authentications rate[2] is related to the accounting start and stop messages as $\xi_{A,i} = p_{a,i}^{-1} \xi_{\text{Start},i} = p_{a,i}^{-1} \xi_{\text{Stop},i} = \lambda_i$. The mean interims rate is the product of the number of interims during each service session and the session arrival rate. Let the session time duration follow a generic distribution $F_S(s)$ with a mean of E_s and a coefficient of variation of $c_s = \frac{\sqrt{Var[S]}}{E_s}$. Let us denote the interim interval as Δ_T and the authorization lifetime as Δ_M. Then, for service i, the number of interims can be obtained by taking the expectation of the floor of the ratio of the duration of the service session and the interim interval Δ_{T_i} as $E[\lfloor \frac{S_i}{\Delta_{T_i}} \rfloor]$. It can be shown that the interim rate from all services is,

$$\xi_{\text{Int}} = \sum_{i \in \mathbb{A}} \lambda_i E[\lfloor \frac{S_i}{\Delta_{T_i}} \rfloor] = \sum_{i \in \mathbb{A}} \lambda_i \sum_{j=1}^{\infty} \bar{F}_{S_i}(j \Delta_{T_i}) \tag{4.2}$$

The mean number of reauthentications can be evaluated similarly to the mean number of interims ξ_R by substituting Δ_{M_i} instead of Δ_{T_i} in (4.2). Thus, the mean signaling rate can be written as,

[2]For brevity, we assume that authentications consist of one exchange, as in 3GPP2 systems. If more than one exchange is needed, such as in WiFi systems which implement EAP authentication schemes, then the authentications and reauthentication rates are multiplied by a constant which reflects their number of messages and processing costs.

$$\xi = \sum_{i \in \mathbb{A}} \lambda_i \left[1 + p_{a_i} \left(2 + \sum_{j=1}^{\infty} \bar{F}_{S_i}(j\Delta_{T_i}) + \sum_{j=1}^{\infty} \bar{F}_{S_i}(j\Delta_{M_i}) \right) \right] \tag{4.3}$$

To get an insight to the general formula in (4.3), let us consider an exemplary case of a single service with an exponentially distributed session duration. It directly follows that (4.3) simplifies to,

$$\xi = \lambda \left[1 + p_a \left(2 + \frac{1}{e^{\frac{\Delta_T}{E_s}} - 1} + \frac{1}{e^{\frac{\Delta_M}{E_s}} - 1} \right) \right] \tag{4.4}$$

which matches the result in (3.9) in Section 3.4. From (4.4), it is clear that there is a non-linear relationship between the interim setting and the mean signaling load. Notice that when $\Delta_T > E_s$, the signaling load barely changes. This is because the mean number of interims per session falls significantly below one (a.k.a, $\frac{1}{e-1} = 0.58$ interim/session).

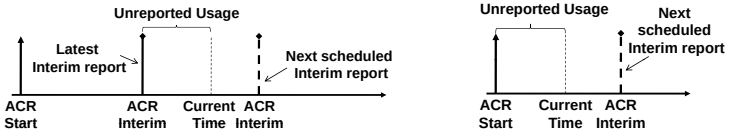
Let us now extend our results to incorporate mobility. In this case, the total signaling rate due to each service is the weighted sum of the signaling load due to its four mobility components denoted as $\xi_i^{(x)}$ and is given as,

$$\xi = \sum_{i \in \mathbb{A}} \sum_{x \in \{F,O,Tr,T\}} \xi_i^{(x)} \tag{4.5}$$

where $\xi_i^{(x)}$ is obtained using (4.3). To obtain an estimate for $\xi_i^{(x)}$ to use in our mechanism and without loss of generality, we assume that the four components of the session holding time, $H_i^{(x)}$, follow the LogNormal distribution as it is widely observed in measurement studies for VoIP and data sessions [95, 96, 98, 99]. Since the complementary distribution for the LogNormal is $\bar{F}_H(h) = \frac{1}{2}\text{erfc}\left(\frac{\ln(kh) - \mu_i^{(x)}}{\sqrt{2\left(\sigma_i^{(x)}\right)^2}} \right)$, then using (4.3) and the results of (3.17) derived in Section 3.4, it follows that,

$$\xi_i^{(x)} = \lambda_i^{(x)} \left[1 + p_{a_i}^{(x)} \left(2 + \frac{1}{2} \sum_{k=1}^{\infty} \bar{F}_H(k\Delta_T) + \frac{1}{2} \sum_{k=1}^{\infty} \bar{F}_H(k\Delta_M) \right) \right] \tag{4.6}$$

where the parameters $\mu_i^{(x)}$ and $\sigma_i^{(x)}$ are given in terms of the mean session holding time and its coefficient of variation as $\mu_i^{(x)} = \ln\left(E_{H_i}^{(x)}\right) - \frac{\left(\sigma_i^{(x)}\right)^2}{2}$, $\left(\sigma_i^{(x)}\right)^2 = \ln\left(\left(c_{H_i}^{(x)}\right)^2 + 1\right)$.



(a) Unreported usage since the latest interim report (b) Unreported usage when no interims are sent yet

Figure 4.8: The unreported usage

4.4.3.2 The Potential Loss

The potential loss L is defined as the unreported usage from impacted services when their serving NAS fails. The potential loss due to a given service i is given as the service consumption since the last interim report or since the service starting instant if no interims were generated yet (see Figure 4.8). For clarity, we first study the potential loss in the absence of mobility and incorporate mobility afterwards. Assuming that the simultaneous failure of multiple NASes is unlikely, the loss due to the failure of a single NAS, L_j, is the sum of the unreported usage from all services belonging to the service set, $\mathbb{N}_j$.

Let us denote the cost of a unit time for service i which belongs to $\mathbb{N}_j$ as C_i and the session duration until the failure moment as $\tilde{S}_i$ with a distribution $\frac{\bar{F}_S(s)}{E_s}$ [94], then using renewal theory, it can be shown that the potential loss due to the impacted NAS j is,

$$L_j = \sum_{i\in\mathbb{N}_j} L_i = \sum_{i\in\mathbb{N}_j} \lambda_i E_{s_i} C_i \overbrace{\left[E\{\tilde{S}_i\} - \Delta_{T_i} E\{\lfloor \frac{\tilde{S}_i}{\Delta_{T_i}} \rfloor\}\right]}^{\varepsilon_i \Delta_{T_i}}$$
$$= \sum_{i\in\mathbb{N}_j} \lambda_i E_{s_i} C_i \varepsilon_i \Delta_{T_i} \leq \sum_{i\in\mathbb{N}_j} \lambda_i E_{s_i} C_i \frac{\Delta_{T_i}}{2} \;,\; \Delta_{T_i} \leq E_{S_i} \tag{4.7}$$

Proof. see the proof in Appendix A in Section A.2.1. □

Let us now briefly discuss the physical interpretation of the potential loss in (4.7). The loss is given as the sum of the products of the losses from all impacted user sessions from all services belonging to the NAS service set (i.e., the $\lambda_i E_{s_i}$ term), the cost of the service per unit time C_i, and the mean unreported usage (i.e., $\varepsilon_i \Delta_{T_i}$). The mean unreported usage is intuitively the difference between the mean *age* of the session time at the moment of failure (i.e., $E\{\tilde{S}_i\}$) and the last interim report of the usage given by

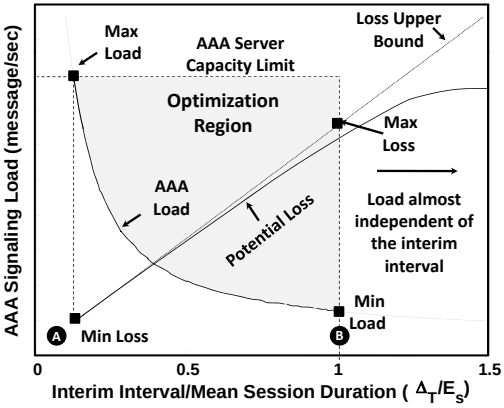


Figure 4.9: The signaling load and potential loss tradeoff (adapted from [145]).

(i.e., $\Delta_{T_i} E\{\lfloor \frac{\tilde{S}_i}{\Delta_{T_i}} \rfloor\}$). For exponentially distributed sessions, similar to (4.4) and due to the memoryless property (i.e., $\tilde{S}_i = S_i$), the mean unreported usage for service i, U_i, is,

$$U_i = E_{s_i} - \Delta_{T_i} E[\lfloor \frac{S_i}{\Delta_{T_i}} \rfloor] = E_{s_i} - \frac{\Delta_{T_i}}{e^{\frac{\Delta_{T_i}}{E_{s_i}}} - 1} \tag{4.8}$$

Observing the limiting behavior of (4.8) as a function of Δ_{T_i}, we notice that as $\Delta_{T_i} \to 0$ then the unreported usage U_i approaches 0 which matches our intuition that continuous interim updates result in no loss at the event of failure. Similarly U_i approaches E_{s_i} as $\Delta_{T_i} \to \infty$. When Δ_{T_i} equals the mean session duration $\Delta_{T_i} = E_{s_i}$, then $U_i \to 0.418\Delta_{T_i} \leq 0.5\Delta_{T_i}$. Hence, in the worst case when the reporting interval equals the mean session duration, the upper bound in (4.7) is only an overestimate by approximately $0.418/0.5 = 16\%$. Thus, in our optimization formulation, we can use the upper bound estimate of the loss which linearizes the loss as a function of the interim interval.

Finally, the potential loss estimate in the presence of mobility is simply obtained by modifying (4.7) by summing the loss due to each mobility component as,

$$L_j = \sum_{i \in \mathbb{N}_j} \sum_{x \in \{F,O,R,T\}} \lambda_i^{(x)} E_{H_i}^{(x)} C_i \frac{\Delta_{T_i}}{2} \tag{4.9}$$

4.4.3.3 The Tradeoff between the Load and the Loss

It is clear from (4.3), (4.9) that there is a tradeoff between the potential loss and the signaling load ξ. This is because if the interim intervals Δ_{T_i} are decreased to reduce

the potential loss in the event of the NAS failure, the corresponding signaling rate ξ increases. To illustrate this behavior, let us for simplicity assume a single service. As shown in Fig. 4.9, the load and the loss are given as functions of the interim interval normalized to the mean session duration. The loss is a linearly increasing function of the interim interval while the load is a nonlinear decreasing function. Notice that when the interim setting is increased beyond the the mean session duration, the AAA signaling load changes very slowly. This is due to the fact that in this case, the session would most likely terminate before any interim messages are sent. On the other hand, significantly reducing the interim values may result in an excessive AAA system overloading resulting in undesired network instabilities (i.e., failovers or redirections). Hence, the desirable optimization region for the interim intervals should be selected such that they neither violate the server capacity, P, nor exceed the mean session duration. In the next section, we design policies to resolve this tradeoff.

4.4.4 The Optimization Policies

In this Section, we propose two optimization policies, i.e., the Constrained Loss Policy (CLP) and the Adaptive Policy with Weight Control (APWC). We also provide a suboptimal method which simplifies the calculation of the CLP by relaxing the load constraint and solve for a bounding loss.

4.4.4.1 Constrained Loss Policy (CLP)

This policy is formulated as a constrained non-linear minimization problem. The objective is to minimize the signaling load ξ from all services subject to two classes of linear constraints: one set limiting the range of the interim intervals for all services within their administrative limits and another limiting the potential loss from each NAS L_i to an upper bound $L_{\max}^{(i)}$. The potential loss for each NAS L_i is calculated based on the interim intervals for services served by it (i.e., $\mathbf{\Delta_T^{(i)}}$). The optimization is subject to the fact that the loss from services served by each NAS i denoted as $L(\mathbf{\Delta_T^{(i)}})$ is below a given maximum $L_{\max}^{(i)}$. The objective function is given by the signaling load $\xi(\mathbf{\Delta_T})$ where $\mathbf{\Delta_T}$ is the union of all interim intervals of all services from all NASes as $\mathbf{\Delta_T} = \cup_{j=1}^{m} \mathbf{\Delta_T^{(j)}}$. Notice that the addition of new NASes due to network expansion or due to failovers simply results in adding new loss and interim range constraints as necessary. If the minimum signaling load exceeds the AAA capacity P, either overload handling mechanisms (such as request redirection [45]) are invoked or the the maximum loss for all NASes is relaxed by a percentage denoted as ε.

As shown in Policy 1, we first check whether a feasible solution exists by comparing the most relaxed settings $\xi(\mathbf{\Delta_T^{\max}})$ to the AAA system capacity P. If the load exceeds the capacity then standard overload handling mechanisms are triggered and

Policy 1 Constrained Loss Policy (CLP)

Input : P, $[\boldsymbol{\Delta}_{\mathbf{T}}^{min},\boldsymbol{\Delta}_{\mathbf{T}}^{max}]$, ε, MaxNumberOfIncreases, $L_{max}^{(i)}$
Output : $\boldsymbol{\Delta}_{\mathbf{T}}$
if $\xi(\boldsymbol{\Delta}_{\mathbf{T}}^{max}) < P$ **then**
  repeat
    IncreaseLmax = false
    if $Loss(\boldsymbol{\Delta}_{\mathbf{T}}^{min}) < L_{max}$ **then**
      Minimize $\xi(\boldsymbol{\Delta}_{\mathbf{T}})$ such that
        $0 < L(\boldsymbol{\Delta}_{\mathbf{T}}^{(\mathbf{1})}) \leq L_{max}^{(1)}$,
        $0 < L(\boldsymbol{\Delta}_{\mathbf{T}}^{(\mathbf{2})}) \leq L_{max}^{(2)}$,
        $\vdots$
        $0 < L(\boldsymbol{\Delta}_{\mathbf{T}}^{(\mathbf{i})}) \leq L_{max}^{(i)}$,
        $\boldsymbol{\Delta}_{\mathbf{T}} \in [\boldsymbol{\Delta}_{\mathbf{T}}^{min},\boldsymbol{\Delta}_{\mathbf{T}}^{max}]$
      if $\xi(\boldsymbol{\Delta}_{\mathbf{T}}) > P$ **then**
        IncreaseLmax = true
        $\forall_k L_k^{(k)} = (1+\varepsilon)L_k^{(k)}$
      end
    else
      IncreaseLmax = true
      $\forall_k L_{max}^{(k)} = (1+\varepsilon)L_{max}^{(k)}$
    end
  until *NumberOfIncreases* > *MaxNumberOfIncreases OR IncreaseLmax* = *false* ;
else
  Trigger overload handling mechanisms;
end

$\boldsymbol{\Delta}_{\mathbf{T}} = \boldsymbol{\Delta}_{\mathbf{T}}^{\max}$ is used. We then check if the maximum allowed loss is possible at the smallest possible reporting intervals. If not possible, we attempt to relax the loss constraints `MaxNumberOfInreases` times before giving up and reporting infeasibility. If all the preconditions are met, the optimization logic is then started by minimizing the load from all services from all NASes (i.e., $\xi(\boldsymbol{\Delta}_{\mathbf{T}})$).

4.4.4.2 The Simplified Constrained Loss Policy (SCLP)[3]

A suboptimal version of the CLP policy can be formulated by solving the constraint equations for each NAS when the loss bound $L_{\max}^{(i)}$ is binding. Although this satisfies the loss requirement, SCLP does not guarantee that the solution results in minimal system load. For the SCLP method, we simply start from the minimal loss at $\boldsymbol{\Delta}_{\mathbf{T}}^{\min}$ and obtain $\boldsymbol{\Delta}_{\mathbf{T}}$ at the NAS loss boundary in one step (see Appendix A.2.2) by moving in the gradient descent direction[4] (i.e., $-\nabla\mathbf{L}$) as $\boldsymbol{\Delta}_{\mathbf{T}} = \boldsymbol{\Delta}_{\mathbf{T}}^{\min} - \alpha\nabla\mathbf{L}$. The constant α represents the

[3] Special thanks to Ankit Singla for his contributions to this policy.
[4] When the gradient for service j is zero, the maximum interim setting for service j is used instead.

magnitude of the movement (see (A.29) in Appendix A.2.2). We then range limit $\mathbf{\Delta_T}$ between $\mathbf{\Delta_T^{min}}$ and $\mathbf{\Delta_T^{max}}$. Afterwards, we check if the load from all NASes is below the system's capacity and if not, we relax the loss limits by moving a small amount ε in the gradient ascent direction as $\mathbf{\Delta_T} = \mathbf{\Delta_T^{min}} + \alpha\nabla\mathbf{L}$ until the capacity limit is satisfied. The SCLP logic is summarized in Policy 2. This suboptimal method, as will be shown in Chapter 5, can be effectively used when an optimizer package is unavailable and when the system's load is not high.

Policy 2 Simplified Constrained Loss Policy (SCLP)

Input : P, $[\mathbf{\Delta_T^{min}}, \mathbf{\Delta_T^{max}}]$, ε, MaxNumberOfIncreases, $L_{max}^{(i)}$
Output : $\mathbf{\Delta_T}$
if $\xi(\mathbf{\Delta_T^{max}}) < P$ **then**
  for each *NAS set* $\mathbb{N}_j$ **do**
    Calculate $\nabla\mathbf{L}$, α using (A.28) and (A.29)
    Calculate $\mathbf{\Delta_T^{(j)}}$ using (A.27)
    range_limit $\left(\mathbf{\Delta_T^{(j)}}\right)$
  end
  while $\xi(\mathbf{\Delta_T}) \leq P$ *AND NumberOfIncreases* $<$ *MaxNumberOfIncreases* **do**
    $\mathbf{\Delta_T} = \mathbf{\Delta_T} + \varepsilon\nabla\mathbf{L}$
    range_limit$(\mathbf{\Delta_T})$
  end
else
  Trigger overload handling mechanisms
end

4.4.5 Adaptive Policy with Weight Control (APWC)

The CLP method requires the setting of loss bounds for NASes which may not be always desirable from operations and management perspective. To address such situations, we propose the APWC policy which does not require the definition of loss bounds on NASes while attempting to optimally minimize the losses using the available capacity and without causing system overload. This policy is formulated as a non-linear minimization problem with a convex objective defined as the sum of the potential loss L from all NASes and a weight (or penalty) function of the signaling load $W(\xi)$ as,

$$L(\mathbf{\Delta_T}) + W(\xi(\mathbf{\Delta_T})) \tag{4.10}$$

The weight function $W(\xi)$ can be any suitable convex function of the signaling load ξ given that it becomes very low when the system utilization ($\rho = \frac{\xi}{P}$) is low and becomes very large when the utilization is high. Here, we use an exponential weight function as,

$$W(\xi) = ae^{\frac{b}{P}\xi} \tag{4.11}$$

where $a = Ke^{-b}$, $b = \frac{\ln(K)}{1-\rho_0}$, and $K = 10L(\mathbf{\Delta_T^{max}})$. In this regard, ρ_0 is a 'knob' parameter which determines the utilization at which the system capacity becomes significant.

Thus, when the system load approaches the capacity limit, the weight function dominates and when the utilization is below ρ_0 the loss dominates. As such, when the load is relatively light the loss is kept as low as possible without causing overload. The constraints include a non-linear convex constraint that the signaling load ξ does not exceed the capacity P and linear constraints on the interim intervals $\mathbf{\Delta_T}$. Policy 3 summarizes the APWC logic.

Policy 3 Adaptive Policy with Weight Control (APWC)

Input : P, $[\mathbf{\Delta}_\mathbf{T}^{min}, \mathbf{\Delta}_\mathbf{T}^{max}]$, $W(\xi)$
Output : $\mathbf{\Delta_T}$
Minimize $L(\mathbf{\Delta_T}) + W(\xi(\mathbf{\Delta_T}))$ such that
$\xi(\mathbf{\Delta_T}) \leq P$,
$\mathbf{\Delta_T} \in [\mathbf{\Delta}_\mathbf{T}^{min}, \mathbf{\Delta}_\mathbf{T}^{max}]$
if n*o solution or loss is too large* **then**
  Trigger overload handling mechanisms;
end

When multiple NASes are reporting to the same AAA server, a representative weighted average for the loss from all NASes is used. In this regard, the NASes with lower potential losses are assigned a lower weight. For instance, consider the case of two NASes with one posing a potential loss of $2,000 while the other posing a risk of losing $20,000 in the event of failure. The arithmetic mean of $11,000 underestimates the real loss of $20,000 if the second NAS fails. To this end, we define weights to the loss from each NAS proportional to its potential loss at a unity interim interval as,

$$\beta_j = \frac{\sum_{\forall i \in \mathbb{N}_j} L_i}{\sum_{\forall i \in \mathbb{A}} L_i} \tag{4.12}$$

Hence, in our two NASes example we have $\beta_1 = \frac{1}{11}$ and $\beta_2 = \frac{10}{11}$. Thus, the loss in (4.10) is given as $L(\mathbf{\Delta_T}) = \sum_{\forall \mathbb{N}_j \in \mathbb{A}} \beta_j \sum_{\forall i \in \mathbb{N}_j} L_i$.

4.4.6 Mechanism Evaluation

In this subsection, we demonstrate the basic operation of our optimization mechanism in fixed network environments and leave the detailed performance evaluation to Chapter 5. To do so, we investigate the mean potential loss and AAA system load (i.e., authentication and accounting requests) in a scenario with two services served by one NAS in a network environment with no mobility (fixed). Services 1 and 2 have mean durations of 5 mins and 15 mins respectively and have equal mean session arrival rates to facilitate comparison. For both services, the mean load varies during the day following a sinusoid with a period of 24 hours and a peak to average ratio of 1.4. The costs for services 1 and 2 are set to 0.1 and 0.4 price units respectively. The tariff for service 2 is halved between 11pm and 6am. For illustration, let us assume that the reduction in the tariff results in doubling the mean session duration from 15 to 30 mins.

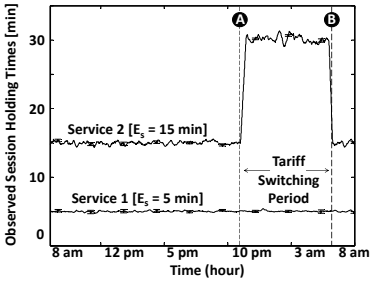
In our evaluation, we implement the proposed mechanism in a JAVA based event driven simulator and link it to MATLAB's Sequential Quadratic Programming method to solve constrained non-linear optimization problems. Our simulation environment consists of several modules for multi-service session generation, network topology and user mobility, and Diameter protocol. The AAA messages are generated according to the AAA standards [40, 41, 45] and according to the accounting model in [44] for mobile networks. Authentications are considered successful by tossing a random variable and comparing to p_a. The service session arrivals are Poissonian and their session durations are generated following Lognormal distributions to match experimental findings for VoIP and wireless data traffic [96, 98, 99]. The optimization logic is only invoked when the statistics change by 5% and when at least a grace period of 75 seconds since the last optimization elapses. The CLP, SCLP, and APWC optimization policies are simulated based on the session statistics using the estimates in (4.5) and (4.9) and are solved using MATLAB Optimization Toolbox. For the APWC policy, we set the knob parameter, ρ_0, in (4.11) to 60%. Finally, since both the AAA signaling load in (4.6) and the potential loss in (4.9) are proportional to the session arrival rates λ_i, all of the results are normalized and given in terms of load (i.e., authentication plus accounting divided by the AAA server capacity) as well as the normalized loss to the target potential loss. Hence, our results apply to arbitrary session loads and AAA system capacities.

In order to assess the benefits of our adaptive scheme, we compare it with three policies with static interim interval settings, to mimic current systems, i.e., Static_Min, Static_Med, and Static_Max. The interim settings for Static_Min are set to 1 min for all services. For Static_Med and Static_Max policies, the interim settings are fixed to half and full mean session durations respectively. For example, for two services of 5 and 15 mins, the corresponding interim settings are [1, 1], [2.5, 7.5], and [5, 15] for the Static_Min, Static_Med, and Static_Max policies respectively. The simulation results are shown in Fig. 4.10 as follows.

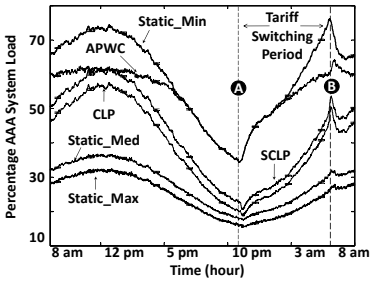
- *The session holding time (Fig.4.10(a)):* The estimated session holding times are equal to the mean session durations for both services due to the absence of mobility. The duration doubles for service 2 in the tariff switching period. The estimate for the arrival rate (not shown) also matches our sinusoidal setting.

- *The system load (Fig.4.10(b)):* As expected, the minimum and maximum loads are achieved by the Static_Max and Static_Min policies respectively and hence the loads of all other policies fall in between. This confirms that the administrative bounds for the interim intervals are respected by our proposed policies. We also observe that for all static policies the load and loss performance clearly follow the sinusoidal session arrival rate which leaves the system load and the potential loss open to the variations in the session statistics (see Fig.4.10(c)).

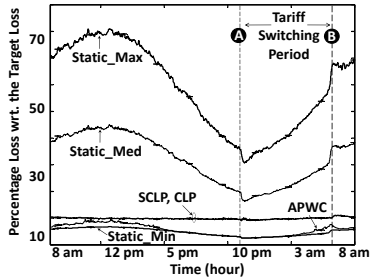
- *The potential loss (Fig.4.10(c)):* For comparison purposes, let us normalize the potential losses from all policies to the target potential loss for the CLP and SCLP mechanisms (i.e., $L_1^{\max}$). As expected, the Static_Min and Static_Max



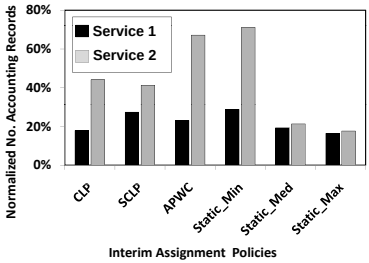
(a) Mean Session Holding Time as Observed by the NAS



(b) AAA System Load



(c) Potential Loss



(d) Number of accounting records per day

Figure 4.10: System's performance in a fixed network environment (adapted from [145]), [tariff switching occurs between 7pm-7am, λ_i=1+0.4sin($\frac{2\pi}{24hr}t$)/s, S/CLP target loss (L_1^{max}) = 400 units, AAA capacity P = 40 req/s, average window sizes = 100, 30 indep. simulation runs, 4 hr warm up period, 95% confidence (change within 3% variation)].

policies set the loss bounds and all policies result in losses that fall in between. For the Static_Med policy, we observe that while halving the interim reporting period for both services only adds 10% extra system load, it potentially results in halving the potential loss. For the APWC policy, we observe that the load curves match the Static_Min policy as long as the load is below our knob setting of 60%. When the load exceeds this setting the loss is increased in favor of lower system load which matches our objective (observe the duration from 8 am to 5 pm in Fig.4.10(b) and Fig.4.10(c)). We also observe in Fig.4.10(c) that both the SCLP and the CLP mechanisms maintain the potential loss target irrespective of the system load with minor 'blips' due to tariff switching. Moreover, the load due to the CLP scheme is lower than that of the SCLP scheme which confirms

the optimality of the CLP scheme.

- *The number of accounting records (Fig.4.10(d)):* For comparison, since Static_Min generates the largest number of interim records, we use the *total* number of accounting records for both services generated by the Static_Min as reference to normalize the number of accounting records from all services generated by all policies. As shown in Fig.4.10(d), for the Static_Max and Static_Med policies, the number of accounting records is almost equal for both services. The slight difference is due to tariff switching which increases the accounting records for Service 2. The accounting records produced by Static_Med slightly exceed those generated by Static_Max due to the lower interim setting of the Static_Med. The accounting records produced by the Static_Min policy primarily reflect the difference in the session durations of both services irrespective of their costs or the AAA system load. The APWC produces less interim records than the Static_Min because it tries to avoid overloading the AAA system by increasing the interim intervals for both services and hence spreading-out the losses. We also observe the similarity of CLP and SCLP in terms of the produced accounting records with the CLP resulting in less total number of interim records. The sub-optimality of the load performance of the SCLP is clear when observing the number of interims produced by service 1 in Fig.4.10(d). Common to all proposed policies (i.e., S/CLP and APWC), service 2 results in more accounting records as it contributes more to the potential loss than service 1 (i.e., $.4 > .1$ price units).

4.4.7 Open Issues

Although we have discussed effective means to efficiently resolve the trade off between the signaling load and the accounting reliability, the following aspects remain as open issues,

- *The cost of the signaling messages*: In our proposed mechanism, we do not directly incorporate the signaling cost of signaling and leave it as a weighting factor for the different types of signaling messages (i.e., authentication, ACR (Start), ACR (Interim), and ACR (Stop)). We follow this approach as the determination of such weights highly depends on the AAA system implementation and deployment environment. Special attention should be followed when determining such costs as in some deployments, AAA systems may incorporate features such as accounting record forwarding/forking to multiple associated network system components such as Wireless Application Protocol (WAP) gateways and content switches, implement special processing for some accounting record types such as updating fraud management systems [49], or proxy the accounting messages to other AAA servers. Such factors can result in variable costs per signaling message type and hence can lead to suboptimal results. Similarly, when the AAA

system is also used to handle authentication and authorization the cost of authentication signaling messages may vary depending on the deployed authentication schemes (e.g., EAP based, PAP or CHAP) and hence careful determination of the signaling message costs is needed per deployment scenario. The automatic determination of the signaling costs per message can be a future avenue for our proposed solution.

- *The impact of load balancing*: Load balancers in the network may result in errors in estimating mobility for users. For instance a user served by some cells in a given region maybe assigned an AGW that usually serves other regions to handle its sessions. When the load balancing rules are deterministic and the user's region is known, the users can be virtually assigned to their respective NAS and the mobility calculation is carried out as we discussed in this section. However, when this is not possible, further investigation is required to determine the impact of load balancing and its frequency on the stability of the estimated mobility statistics (i.e., avoiding very high coefficients of variation).

- *Incorporating Prepaid and Converged Billing Systems*: First, in prepaid billing systems, the AAA system interacts with the prepaid server as we discussed in Chapter 2. This results in an additional signaling load that requires to be estimated as well. In addition, prepaid users may also be dropped during service if they deplete their quota. Although this aspect may not be significantly impacting, it is still instructive to investigate its effect on the mechanism. Second, in converged billing systems, accounting records should be processed in realtime by the billing system. Thus, a future avenue for this research is to incorporate the performance of BSS components as part of the optimization polices. In addition, further work is also needed to incorporate pricing tools as in [54] to dynamically provide/update service costs for our mechanism.

- *When accounting is enabled on a per cell basis*: In this case, the signaling load estimation block needs to be further extended to consider the cellular channel holding time at each base station and mobility patterns between base stations, in order to estimate the aggregate NAS signaling load.

Now that we have investigated AAA protocol optimizations, we proceed to discuss potential AAA applications in cellular backhaul and in layer 2 optical communications.

4.5 AAA Applications in Cellular Backhaul over WMN Deployments

In this section, we propose an architecture which facilitates accounting for wireless cellular backhaul deployments over Wireless Mesh Networks (WMN). Our application

is motivated by the fact that wireless mesh networks have been recently proposed as a possible cellular backhaul transport media, not only for their significant cost savings but also for their scale, flexibility and resilience. However, wireless mesh backhauls pose many technical challenges including timing synchronization for GSM networks, bandwidth reservation and control techniques, as well as accounting mechanisms by mesh operators. In depth description of cellular backhaul issues and future trends is available in our article in [149].

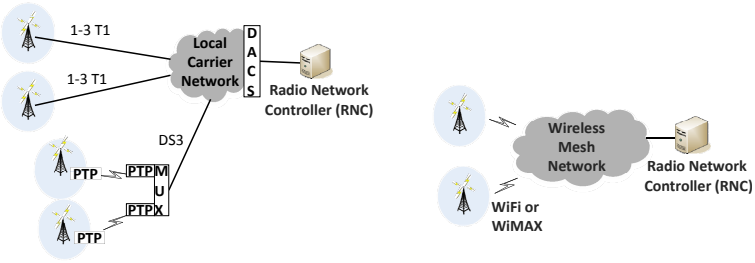
In this section, we discuss the business case in which wireless mesh operators offer their services to cellular operators for backhaul services. We propose the first billing architecture for cellular backhaul applications over wireless mesh networks and analyze its scalability. While this is only the first step to address the generic area of billing for multi-service wireless mesh networks, when applied to cellular backhaul it poses few practical yet new challenges in the context of accounting signaling. First, adding or releasing backhaul bandwidth chunks directly reflects on the signaling for billing updates. Second, the performance of every billing scheme highly relates to the dynamic bandwidth reservation mechanisms at the base station. It is very critical that the billing signaling traffic resulting from bandwidth reservations is minimal to ensure scalability for the billing architecture. Since the admitted rates may vary with the implementation of the reservation mechanism, we study a relatively poor implementation and establish upper bounds on billing signaling rates in response to the used reservation mechanism.

In the following subsections, we provide a short background relevant to cellular backhaul and describe our billing architecture as well as the signaling mechanisms. At the end of the section, we evaluate our mechanism using Markov chains.

4.5.1 Background

With the tremendous success of cellular telephony and the high demand on future media services, a scalable, flexible, and cost effective backhaul transport technology is becoming an absolute necessity. Current backhaul technologies from the base station (BTS) to the radio network controller (RNC) are primarily based on T1/E1 or microwave link technologies (see Fig. 4.11(a)) [150]. It is widely recognized that the current technologies are difficult and expensive to deploy; for instance, line-of-sight reception is a requirement for proper operation that may not always be feasible. Furthermore, current technologies may not always offer cellular operators with the chance to incrementally modify the backhaul and upgrade it with cost-reducing technologies. Put simply, the deployment of the BTS in its optimal location and with optimal performance may not be always possible.

To address these issues, Wireless Mesh Networks (WMN) have been recently proposed as backhaul media for cellular systems as shown in Fig.4.11(b). Utilizing WMN for backhaul communications presents a promising avenue for backhaul cost savings and easier BTS site deployment by relieving the requirements of the availability of T1 con-



(a) Current cellular backhaul technologies

(b) Wireless mesh networks for cellular backhaul

Figure 4.11: Current and emerging cellular backhaul technologies.

nections or Line of Sight (LOS) connectivity for microwave links. Moreover, WMNs also offer flexibility and resiliency, where, for example, a BTS can failover to different RNCs in cases where an RNC fails. The fact that a few commercial wireless mesh backhaul deployments already exist today confirms their wider considerations [16].

Using WMNs for backhaul communications poses a significant challenge as various technical issues must be considered including timing and synchronization for GSM systems, routing and scheduling, security, resource reservation, and billing. For example, in [151], the authors present BTS timing synchronization challenges for replacing T1 links for GSM/UMTS systems with wireless links and they proposed solutions using GPS and the IEEE 1588 PTP protocol to resolve such issues. In [152], routing and scheduling issues were jointly treated to optimally route packets over mesh networks in WiMAX based deployments. In [153], overhead analysis and enhancements for deploying point-to-point WiMAX backhaul links were proposed. In [154], the authors proposed a reservation protocol (called DARE) similar to ReSerVation Protocol (RSVP) for WiFi mesh networks to address the issue of the multi-hop bandwidth management.

4.5.2 System Design

In this subsection, we address two important new proposals: (i) the billing architecture and signaling for wireless mesh backhaul, and (ii) the bandwidth reservation mechanism and its relationship with (i). We assume that the issues of security, routing stability, as well as interference mitigation are properly handled (e.g., see [152, 154]).

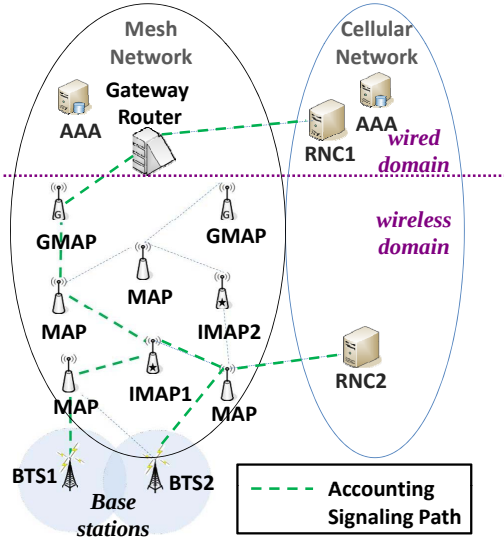


Figure 4.12: Billing architecture with wireless mesh backhaul (adapted from [20]) [MAP: Mesh Access Point, GMAP: Gateway MAP, IMAP: Intermediate MAP, BTS: Base Transceiver Station, RNC: Radio Network Controller, AAA: Authentication, Authorization, and Accounting].

4.5.2.1 Billing Architecture Overview

Fig. 4.12 shows our proposed architecture. A cellular operator connects the base stations (BTS) to RNC1 and RNC2 through the wireless mesh (dashed lines). The mesh network consists of multiple Mesh Access Point (MAP)s and Gateway Mesh Access Point (GMAP)s. The GMAPs connect the wireless mesh network to the wired IP backbone (i.e., Internet). The BTS and RNC elements may either communicate through the wired networks or over the wireless mesh. RNC1 is best reached through the GMAPs by the wired network domain, while RNC2 is best reached through the wireless mesh. For BTSes connecting to RNC1, the gateway router acts as a network access server and generates billing records towards the acAAA system. On the other hand, for communication between BTS2 and RNC2, the edge MAPs monitor usage and report it to an Intermediate Mesh Access Point (IMAP). The IMAP then forwards the usage (billing) records to the mesh operator's AAA server. IMAPs act as collection points and can be viewed as mini-accounting servers. Notice that multiple IMAPs maybe crossed before reaching the AAA server. For example (not shown in Fig. 4.12), IMAP1 may forward to IMAP2 and then to the AAA server.

In our architecture, IMAPs are special MAPs with enhanced physical security. IMAPs include accounting storage capabilities which allow relaying accounting records to the mesh operator's AAA server. IMAPs announce themselves using simple hop-count limited broadcasting. Thus, for a grid layout mesh, geographically close MAPs with client BTSes within the mesh are informed of the existence of the IMAP. To guarantee the operation of the billing service, static last resort IMAPs may be configured in the MAPs. Thus, MAPs with clients such as BTS2 and RNC2 in Fig.4.12 report accounting to IMAP2 and the accounting traffic would eventually be routed to the mesh operator's AAA. Notice that IMAP1 acts also as a simple MAP to connect BTS2 to RNC2.

For very large mesh networks, it might be too costly to have the MAPs implement application layer accounting protocols such as RADIUS or Diameter [41, 45]. Furthermore, the IMAP may not be able to manage a large number of associations from edge MAPs. In this case, MAPs may simply elect a root MAP within a certain distance and use light-weight messages to update the root node with the current usage statistics relevant to billing. The elected root nodes would report usage to the IMAP and thus limit the implementation of the RADIUS/Diameter to the root nodes. This approach in essence distributes the network access server (NAS) function [100, 155]. To simplify the discussion, we will assume that all MAPs report usage directly to the IMAPs.

In our design, the accounting process of RADIUS and Diameter including start, interim, and stop messages is followed. Notice that interim reporting is used to avoid monitory losses in case of MAP failures (i.e, the metering MAPs). In all cases, the accounting interim interval can be statically configured or can be passed as an authorization parameter to each MAP. Furthermore, interims are sent in response to a bandwidth reservation process. It is therefore very important to relate the bandwidth reservation dynamics to the billing signaling rates. As we will see next subsection, the interim interval only controls the minimum billing traffic rate to the IMAP node.

Finally, when the two communicating nodes lie within the mesh network (e.g., a BTS and a RNC), we assume that a simple IMAP initiated Diameter authorization message can be sent to one of the two edge MAPs to disable its accounting messages. This can be communicated using a Vendor Specific Attribute (VSA) defined for this purpose [45]. For instance, only the edge MAP for RNC2 is allowed to report traffic to the IMAP while the billing reports from BTS2's MAP are disabled.

4.5.2.2 Bandwidth Reservation Scheme

The BTS load varies depending on the time of the day, specific events, etc. The highest efficiency can be achieved by reserving bandwidth at the backhaul when the users are admitted to the cellular network. However, the post dial delay (a QoS parameter) [19] maybe severely impaired depending on the reservation delay over the mesh. Furthermore, excessive reservation and billing signaling overhead will be required. On the other hand, the minimal signaling overhead and the minimal post dial delay are achieved

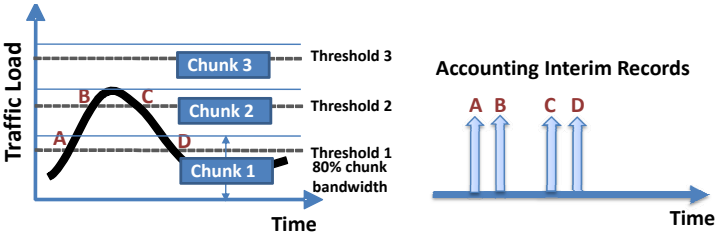


Figure 4.13: Threshold based reservation scheme (adapted from [20]).

when the full bandwidth is reserved by the base station when it initially attaches to the mesh network. With such a tradeoff in mind, we propose that a threshold based bandwidth reservation mechanism be used at the BTS level. This is illustrated in Fig. 4.13. The BTS initially reserves a predefined amount of bandwidth to handle a subset of users (i.e., chunk 1). The reservation can be performed using mesh reservation protocols such as DARE [154]. Reservation requests are normally made by the BTS when the measured traffic load crosses a threshold. For example, when the traffic load reaches point A in Fig. 4.13, the BTS requests the reservation of chunk 2 from the mesh network. To avoid ping-pong signaling, hysteresis can be used such that thresholds in the increasing direction are different from the ones in the decreasing direction. For example, Threshold 1 in Fig. 4.13 can be set at 80% of chunk 1's capacity to claim chunk 2, while it can be set at 70% of chunk 1's capacity for the decreasing traffic to release chunk 2.

Upon reservation, the MAP immediately sends an accounting interim to the IMAP reflecting the reservation update. The periodic interims may or may not be affected by the transmission of the last interim record. In one implementation, the next periodic interim can be sent one interim interval after the reservation triggered interim. Alternatively, the periodic interim can be sent at the original schedule and is unaffected by the transmission of the reservation triggered interim. The choice depends on the post processing complexity. Figure 4.14 illustrates our BTS reservation scheme and its relationship with the billing signaling rate. For the sake of our discussion, traffic could be simply viewed as the number of the admitted connections (e.g., guaranteed voice and data services) or by using more sophisticated traffic measurement schemes for example by employing a moving average window to measure the actual traffic rate for different service classes and reserve bandwidth according to the spare capacity. Detailed traffic modeling is out of our scope.

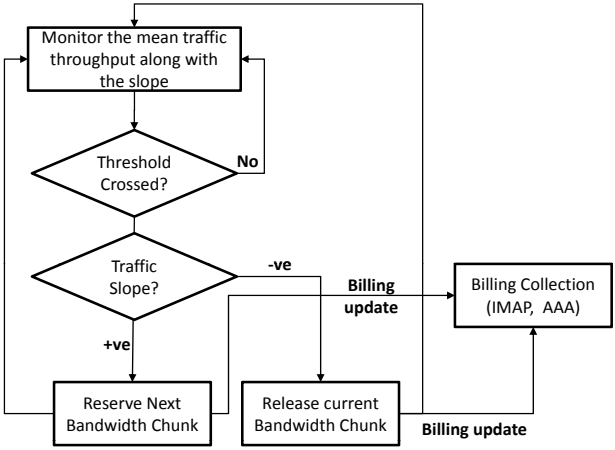


Figure 4.14: Bandwidth reservation and billing flow diagram (adapted from [20]).

4.5.3 Modeling the Accounting Signaling Rate

We now show that our proposed threshold based mechanism combined with the proposed billing architecture can scale without significant signaling overhead. For simplicity, we only consider VoIP service class. Voice is commonly modeled by an ON/OFF source where it transmits a constant rate r_{ON} bps while in the ON state and r_{OFF} bps while in the OFF state.

4.5.3.1 Assumptions

- A buffer-less system serving a maximum of C VoIP flows.
- Flows arrive according to a Poisson process at rate of λ
- The duration of the flows is exponentially distributed with a mean of $\frac{1}{\nu}$.
- A constant rate r_{ON} bps is reserved for each flow.
- No hysteresis is assumed.

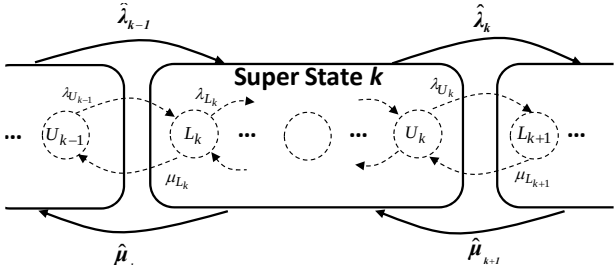


Figure 4.15: State aggregation based on thresholds (adapted from [20]).

4.5.3.2 Analysis

The dynamics of the system can be described by a simple one dimensional Markov process (n), where n is the number of flows in progress. The associated Markov chain is shown by the dashed states in Figure 4.15 and has the transition rates,

$$\lambda_i = \lambda \quad , \quad \mu_i = i\nu \tag{4.13}$$

Its stationary state probabilities are given by the Erlang distribution as,

$$P_i = \frac{A^i/i!}{\sum_{j=0}^{C} A^j/j!} \quad , \quad 0 \le i \le C \tag{4.14}$$

where $A = \lambda/\nu$ is the offered traffic. While in state n, on the backhaul, a reservation of $R(n) = n r_{ON}$ is required for the admitted flows. In the threshold reservation method (with no hysteresis), the reservation and the billing signaling are triggered only when the traffic crosses a threshold. This is modeled by a set of K+1 thresholds: $T = \{T_0, T_1, ..., T_K\}$, where $T_0 = 0$, which are used in such a way, that all states of the Markov chain with reservations $T_k < R(L_k) < ... < R(U_k) \le T_{k+1}, k = 0, 1, ..., K-1$ are aggregated together forming a super state k, see Fig. 4.15. L_k and U_k are the "micro" states at the boundaries of each aggregation. Based on this aggregate model, the mean signaling rate is the rate of leaving the super states.

The probability of being in a super state k, denoted as p_i, formed by the micro states ranging from L_k to U_k is given by the sum of their probabilities as,

$$\pi_k = \sum_{i=L_k}^{U_k} p_i \quad ; \quad k = 0, \ldots, K \tag{4.15}$$

The non-zero elements in the generator matrix $\mathbf{Q}'$ for the super states are given as,

$$\begin{aligned} \grave{q}_{k,k+1} &= \hat{\lambda}_k = \lambda \frac{p_{U_k}}{\pi_k} \quad , \quad k = 0, \ldots, K-1 \\ \grave{q}_{k,k-1} &= \hat{\mu}_k = \mu_{Lk} \frac{p_{L_k}}{\pi_k} \quad , \quad k = 1, \ldots, K \\ \grave{q}_{k,k} &= -\left(\hat{\lambda}_k + \hat{\mu}_k\right) \end{aligned} \tag{4.16}$$

Although the stationary probabilities of the aggregate process satisfy the global balance equations $\pi \cdot (\mathbf{Q}')^T = \mathbf{0}$, where T denotes the transpose, it is no more a simple Markov chain, because the distribution of time spent in a state has now a phase type distribution. Since the mean time spent in any super-state k is given by $(-\grave{q}_{kk})^{-1}$, the mean signaling rate can be obtained as the sum of the products of the state probabilities and their corresponding departure rates as,

$$E\left[\xi_{\text{MESH}}\right] = -\sum_{k=0}^{K} \pi_k \grave{q}_{k,k} = -\pi \, diag\left(\mathbf{Q}'\right)^T \tag{4.17}$$

To calculate the accounting traffic rate, we only consider the interim traffic rate as the BTS will be always connected to the mesh network and is unlikely to disassociate from the network frequently. In other words, the accounting start and stop messages due to BTSes associating and dissociating with the mesh network are insignificant and can be ignored. Assuming a fixed accounting time interval of Δ_T, then if the triggered interim does not time-shift the following interim, the mean accounting traffic rate is given as,

$$E[B_{ns}] = E\left[\xi\right] + \frac{1}{\Delta_T} \tag{4.18}$$

The other case where the interim shifts with the last update results in less signaling than in (4.18). Since in many cases the interim-interval (Δ_T) is set to values in the order of minutes, it can be shown that the mean billing traffic rate for relatively large values of Δ_T (i.e., compared to $E\left[\xi\right]$) is given as $E[\xi] \approx E[B_{ns}]$.

4.6 Simulation and Numerical Results

For practical relevance, let us analyze the case where an EVDO operator connects the base stations to RNCs through a mesh network in an architecture shown in Fig. 4.16. The system parameters are summarized in Table 4.3. For simplicity, let us also assume that the BTSes are omni-directional and only support VoIP traffic. For comparison purposes, we define the per-flow and the flat reservation as extreme reservation schemes.

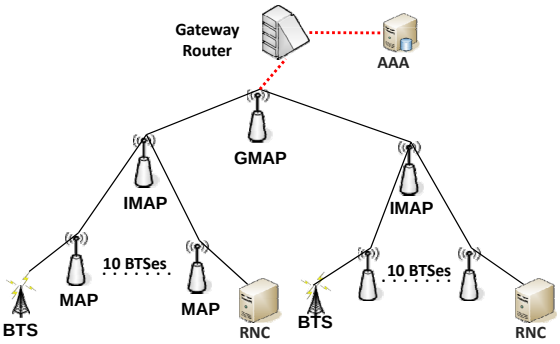


Figure 4.16: A Sample topology for cellular backhaul over wireless mesh (adapted from [20]).

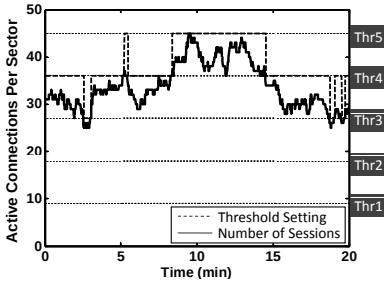
Per-flow means reserving bandwidth for each incoming user resulting in maximum signaling load, while flat refers to a reservation at a full BTS bandwidth which results in almost no signaling load excluding the interims. In the simulations, we generate user arrivals according to Poisson distribution and call durations according to a negative exponential distributions. With each event, we check whether the thresholds are crossed and if so a billing message is generated towards the IMAP, according to the process illustrated in Fig. 4.14. Notice that our mechanism is triggered based on the number of requested connections and without applying any hysteresis. This is to illustrate that our reservation method scales even under poor reservation implementation assumptions. We also compare our simulation results to the analytical results in (4.18).

Table 4.3: Simulation parameters.

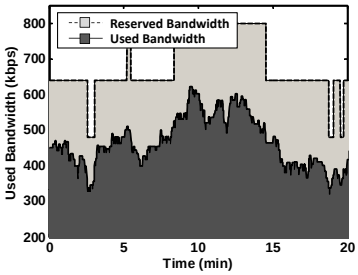
VoIP Sources	A talk spurt of 5 sec, voice activity factor = 43.5%
EVRC Codec	21.45 Bytes/20 ms for active sources [full rate 22 Bytes and 1/2 rate 10 Bytes with probabilities (41.5% and 2%)] 2 Bytes/20 ms for inactive sources [i.e., 1/8 rate with 56.5% liklihood]
Overhead	Robust Header Compression (ROHC) = 2 Bytes EVDO MAC Layer = 17 Bytes
BTS Capacity	35 Erlangs [156] per sector, 3 sectors per BTS, 2% blocking, 45 connections/sector

Figure 4.17 illustrates our numerical results. Since our mechanism is triggered by the number of connections served by base stations, we first show the number of active connections as well as the reserved and used bandwidth per base station. In Fig.4.17(a), we show the number of active connections over an exemplary 20 min period and how crossing thresholds results in AAA signaling. In Fig.4.17(b), we show how the re-

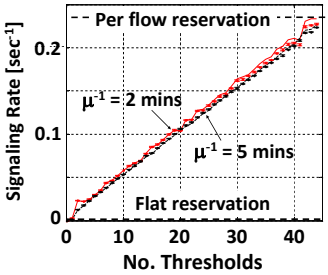
served bandwidth dynamically changes as function of the used bandwidth. Although this scheme is relatively inefficient, it performs better than static approaches with fixed bandwidth (e.g., T1 1.5 Mbps bandwidth).



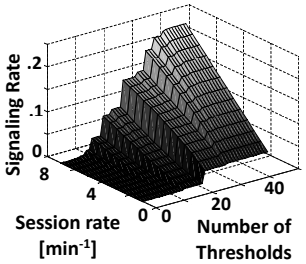
(a) Number of active connections



(b) Used bandwidth per base station (kbps)



(c) Signaling rate as function of the number of thresholds



(d) Signaling Rate as function of the number of thresholds and the session arrival rate

Figure 4.17: Operation of the AAA backhaul application (adapted and modified from [20]).

Fig.4.17(c) shows the simulation as well as the numerical results for the cases of 14 Erlangs (μ^{-1} = 2 mins) and 35 Erlangs (μ^{-1} = 5 mins). The simulation results match the analytical model very well, within 95% confidence. Again, when the number of thresholds reaches 44 (i.e., one threshold per user), the threshold based mechanism becomes a per flow reservation scheme. It is easy to show that, if we neglect blocking, the signaling rate of the per flow is approximately 2λ (2*7/60 = 0.23, in our case). In practice, the number of thresholds is expected to be set to values from 1-5 thresholds to reduce the signaling overhead as shown in Table 4.4. From Table 4.4 and Fig.4.17(c), we notice that for high number of thresholds (> 5) the difference in the signaling rates is almost insensitive to the call duration.

Table 4.4: Some signaling rates for a set of practical thresholds [20] [Arrival rate = 7 calls/min].

No. thresholds	**1**		**3**		**5**	
μ^{-1} **[mins]**	**2**	**5**	**2**	**5**	**2**	**5**
Signaling rate/min	0.17	0.06	1.34	1.05	1.80	1.72

Finally, in Fig.4.17(d) we see the signaling rate as a function of the arrival rate and the number of thresholds. The thresholds are uniformly sized and are set at the chunk maximum capacity (i.e., 100% level). The quantization effect due to the thresholds in the Markov chain model is the primary reason behind the stair-like shape of the signaling rate. Note that when the number of thresholds are set to 44 in this example (i.e., reflecting the maximum BTS capacity of 45 connections), the thresholds based method becomes a per arrival reservation scheme resulting in the highest signaling rate.

We conclude our results by an assessment of the IMAP processing and storage requirements. From Table 4.4, we see that for a practical threshold setting of 5 that the signaling load is 1.8 updates/min/BTS. From the topology in Fig. 4.16, the aggregate rate to each IMAP (i.e., from 10 BTSs) is 0.3 updates/sec. Such values have an important practical relevance as they can be easily handled with current hardware. The storage requirements can be estimated by assuming a typical Diameter message size of 2 KB. If each IMAP is required to keep billing records for 1 day (a realistic assumption), then the required storage capacity is approximately 52 MB. Such system parameters can be easily met with today's storage technologies. Notice also that such an estimate may be even an overestimate as the load may considerably drop during certain hours and hence less capacity may be consumed. In a similar fashion, it can be shown that the per flow reservation requires IMAPs to support 2.3 updates/sec and 397 MB storage. Such rates may not scale well for larger topologies, with larger number of IMAPs. In summary, the results show a remarkably low accounting signaling traffic generated by the examined reservation scheme, which scales linearly with the number of base stations.

4.6.1 Open Issues

Further research is still required towards reaching a comprehensive billing solution for wireless cellular backhaul deployments. The following are some areas to consider,

- *The impact of the the traffic variability at base stations*: In this section, we majorally considered connection oriented traffic such as voice as a major source for base station traffic in order to demonstrate the feasibility of our proposed mechanism. However, future research is needed to determine the impact of data traffic variability on the number of reservation thresholds by the base stations as this depends on the deployed base station technologies.

- *The number and the location of the IMAPs*: Further investigation is needed to determine the optimal location for the IMAP nodes as well as their numbers. Ideally, the number of IMAPs should be kept to a minimum to avoid higher management costs and to minimize the security risks.

- *Network management aspects*: Control plane signaling mechanisms are needed to select the metering MAPs and configure the accounting reporting parameters at the MAPs and IMAPs. In addition, control plane signaling is needed to set the thresholds for the base stations. Consideration of MAC layer signaling versus application layer signaling should be also evaluated.

- *Accounting protocol message format*: Further investigation is required to design the accounting message format while considering airlink load, processing capabilities of mesh nodes, and storage of the IMAPs and their numbers.

4.7 Authentication in Multi-Domain Optical Networks[5]

Another interesting application for AAA signaling arises in multi-domain layer 2 optical network environments. In this case, a layer 2 data path between the source and the destination may span over one or more transit optical domains as shown in Fig. 4.18. Paths are computed using the Path Computation Element (PCE) framework which was recently proposed as potential solution for inter-domain service provisioning with constraint-based path computation in carrier grade Ethernet and Multi-Protocol Label Switching (MPLS) networks [157, 158]. Within the PCE framework, the local Traffic Engineering Database (TED) is used to compute optimal paths inside a single domain and inter-domain path computation is facilitated by sequentially computing a virtual shortest path tree from the destination to the source domain. The PCE framework allows requesting a path from non-neighboring domains where the source domain requests path segments in remote domains to setup multi-domain connections with QoS guarantees. In real deployments, however, the current PCE framework [159, 160] does not address issues of Authentication and Authorization (AA) in business models which encompass neighboring and non-neighboring domains alike [15]. Furthermore, given that the source domain explicitly requests QoS levels in remote domains based on a business relationship, the source domain must also be capable of billing and auditing of the connection in these domains.

The direct application of the current AAA models [40, 41, 45, 46] does not work due to the fact that a data connection in our context may be served by multiple transit carriers which need to authorize its establishment and meter its usage. In fact, this can be conceptually viewed as a generalization for the roaming scenarios discussed in Chapter 3, where the data path is generally maintained by the visited network or extended to the home network (see Section 3.6). This generalization necessitates extending the

[5]This section is a result of collaboration with Mohit Chamania and Silvana Greco

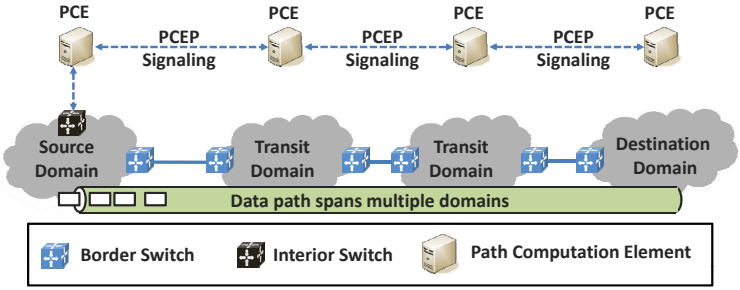


Figure 4.18: The Policy Computation Element (PCE) usage in multi-domain environment [Acronyms, PCEP: Path Computation Element Protocol]

current authentication and authorization (AA) schemes to suite multiple carriers as well as redefine the accounting signaling procedures where multiple networks implement the network access server (NAS) metering functionality.

In this section, we present a new signaling framework that extends the PCE functionality to support authentication and authorization mechanisms on a per connection basis in PCE enabled inter-carrier networks. The proposed authentication and authorization (AA) signaling extensions facilitate authenticated exchange of path computation signaling among domains, allow authorization for the requested QoS levels, and securely associates resource reservations with the computed paths. In addition, we propose the signaling for accounting to enable charging for the service usage by distributing parameters needed to produce accounting messages by the participating domains.

4.7.1 Background

The PCE framework allows networks to interact with non-neighboring domains and *compute* data paths that optimally satisfy the required QoS requirements. The computed paths are then used by reservation or connection (path) setup protocols (e.g., RSVP) to reserve resources within the participating domains. From a system perspective, the PCEs are used along with a local traffic engineering database (TED) to compute optimal intra-domain paths. For inter-domain communication, PCE's of different domains interact with each other using the PCE communication protocol (PCEP) [160] to compute inter-domain paths. From a signaling protocol perspective, a PCE path computation request is sent from the source to the destination domain's PCE along the desired domain chain. The destination domain's PCE creates a Virtual Shortest Path Tree (VSPT) description which consists of path segments from the destination switch (node) to the relevant border switch. The VSPT description along with path keys are sent to the

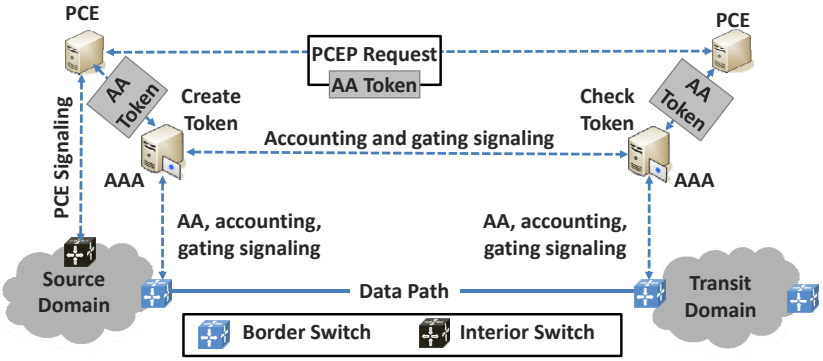


Figure 4.19: Extended PCE framework with AAA.

upstream transit domains in the PCE response messages. Path keys [161] are used to request the setup of certain computed paths within transit domains during the path setup phase without sharing topological information between carriers. Afterward, each subsequent domain adds the VSPT and path keys to PCE response message, which upon reaching the source PCE, allows the computation of the optimal inter-domain path from the source to the destination node using the full VSPT information from all domains.

In the next subsections, we propose a new Diameter application which allows authenticating path computation (PCEP) requests and securely links the computation to the path setup phase in RSVP. In addition, we utilize the Diameter gate control signaling, which we mentioned in Section 2.4.3, to control the data flow within the established paths. Finally, we show how the accounting process functions when multiple NASes belonging to the transit and destination domains are reporting accounting information to the source domain. It should be noted that this section is a major extension to the work in [15] which proposed an AA mechanism to authenticate PCE computation requests, but left path setup, control, and accounting open for future research endeavors.

4.7.2 Introducing AAA to the PCE framework

In our architecture, we assume that the source domain has business relationships with remote (non-neighboring) domains which allow the PCE framework to establish QoS paths. As shown in Fig. 4.19, a Diameter AAA server is added to the existing PCE architecture. As it can be seen in Fig. 4.19, both the PCE and the border nodes are equipped with a Diameter client and interact with the AAA system. PCE and path reservation messages (e.g., RSVP) are modified to include additional information for authentication and authorization of the signaling messages. Upon the arrival of a PCE/RSVP

message, the authentication and authorization information is extracted by the Diameter client to perform these functions with the local AAA server. The authentication and authorization information is typically inserted in the form of tokens. The border nodes are also equipped with metering and gating functions to support inter-carrier accounting and data plane forwarding control. The accounting application running at the border nodes is essential for auditing and billing. As the source has business relationships with remote domains, the source AAA server has peering agreements with all the remote AAA servers. Therefore, accounting messages generated in remote domains are proxied via their AAA servers to the source's AAA server. In a similar fashion, gate control commands sent by the source domain are first forwarded to the domains' AAA servers, which proxy the message to the corresponding border nodes.

The signaling flow is shown in Fig.4.20. We see that the AAA signaling consists of authentication and authorization (AA) phase followed by an accounting phase for established connections. Within the AA phase, we identify three distinctive phases, i.e., path computation, resource reservation, and gate control. The accounting phase is used for established connections only. Accounting messages are primarily used for billing and auditing purposes, but they can also be useful for monitoring or failure localization messaging.

In short words, the PCE path computation request carries a unique connection identifier, the requested QoS levels, and a digitally signed token generated by the source which all transit domains can verify. The PCEP request also carries the timeout information in which the source suggests a validity period for the requested path to the transit domains. Every PCE request is followed by the PCE response which carries parameters generated by each transit domain including computed path tokens pertaining to the domains' computed path keys, the actual path key timeout periods, and unique Reservation IDs that will be used in the path reservation phase. When the source domain decides to perform the path setup, the corresponding reservation signaling, typically RSVP, carries a resource allocation token which contains the Reservation IDs for the computed path as well as parameters relevant to the accounting process for the connection, such as the interim intervals and the Multi-Acct-Session-ID. The latter is used to correlate accounting information from all transit domains to the source to a unique accounting session. Once the path is reserved, data plane forwarding is activated via the AAA infrastructure using path activation or gate control mechanisms and the accounting process is initiated. Afterwards, each transit domain reports usage to the source using accounting messages which have session identifiers unique to the transit domain and a single Multi-Acct-Session-ID that is unique to the whole connection. Although this is similar in principle to the 3GPP2-Correlation-ID or 3GPP-Charging-ID which relates accounting records from AGWs when users move in their respective areas, this is different in the way that accounting signaling from all participating domains occurs concurrently while in cellular networks it happens sequentially. Let us now describe the details of the signaling flow for the proposed mechanism as depicted in Fig. 4.20.

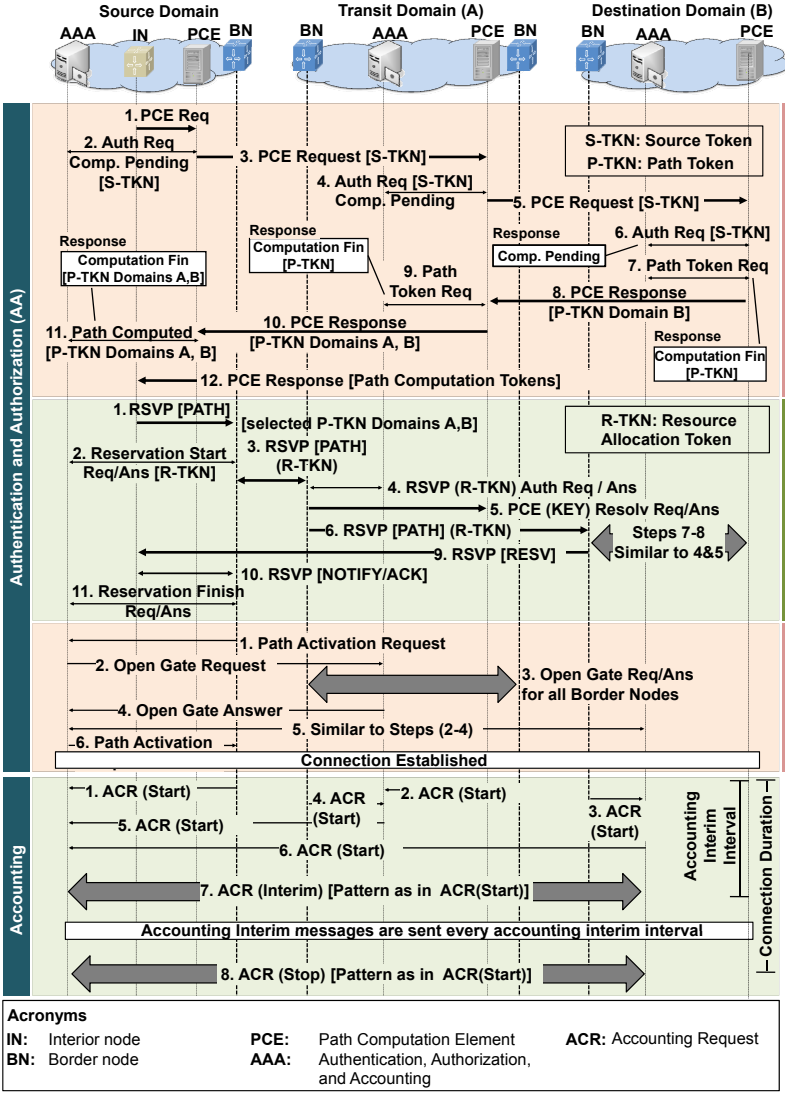


Figure 4.20: Signaling for Path computation setup and accounting in multi-domain systems [Accounting answer (ACA) messages are not shown for clarity].

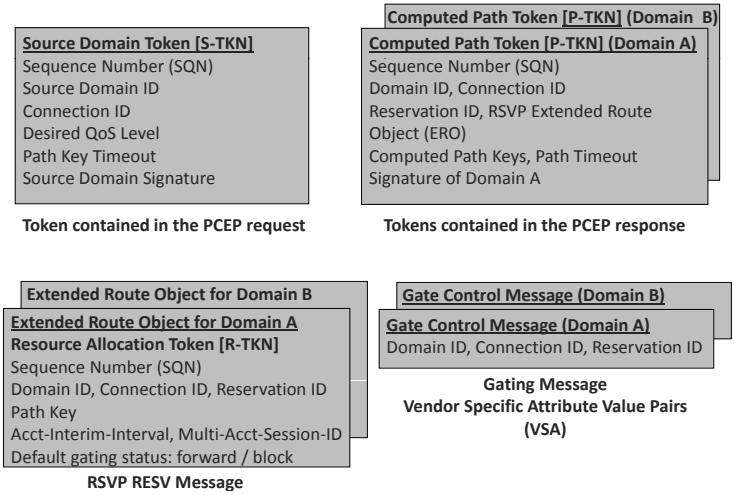


Figure 4.21: Information exchanged in PCEP, RSVP, and Diameter gate control messages.

4.7.2.1 AA for Path Computation

After the source node sends a path setup request to the domain's PCE(1), the PCE determines the domain chain for the path computation request, and sends a *Authentication Request*(2) to the AAA server to create a digitally signed source token (S-TKN). The S-TKN is included in the computation request and is used by the transit domains to authorize the request. The token includes the desired QoS levels from each transit domain, source domain ID, a randomly generated connection identifier which serves also as a nonce for AA, a continuous sequence number, and the suggested timeout for the computed path keys as indicated in Figure 4.21. The sequence number is used to prevent replay attacks, wherein a malicious entity may re-transmit a valid PCE request at a later time. The token is returned in a *Computation Pending* message and the connection request including the source token is then forwarded to the transit domains(3). In step (4), the remote PCE consults its AAA system to authenticate the source token and to authorize the requested QoS levels via the Auth Req. The AAA server verifies the source token and authorizes the requested QoS levels according to the existing service level specification (SLS) and responds with a Computation Pending message. Similar messaging is performed in all transit domains until the destination domain (i.e., step 6 in our example).

Upon successful authorization of the PCE request, the destination domain's PCE computes the path segments according to the Backward-Recursive PCE-Based Computation

(BRPC) protocol, and generates a path key for each computed path. The PCE then requests a computed path token from its AAA system by sending a Path Token Req message(7). The AAA system sets the path key timeout value and generates a digitally signed path token(P-TKN) as shown in Fig. 4.21. The computed token is sent to the PCE via the Computation Fin message, indicating the termination of the AAA session in the local domain, and the received token is inserted into the PCE response message(8) along with the QoS parameters of the computed paths. This procedure is repeated in each of the transit domains (9-10) until the PCE response reaches the source domain. In the source domain, the PCE sends all the computed path tokens to the AAA server for verification via the Path Computed message(11). The AAA verifies and signs the whole set of the transit tokens and returns them to requesting PCE via the Computation Fin message. Once the PCE verifies the response, it returns the computed path and the signed computed path tokens to the source node in the PCE response message(12).

4.7.2.2 AA for Path Setup

When the source domain decides to setup the path, RSVP signaling is initialized with the RSVP [PATH] message(1) which includes the signed computed path tokens. When the RSVP [PATH] message arrives at the egress node in the source domain, the border node extracts the computed path tokens from the RSVP message and sends them to the AAA server via the Reservation Start Req message(2). The AAA server verifies the source domain's signature. It then generates and signs resource allocation tokens (R-TKN) which contain reservation identifiers and accounting parameters (see Fig. 4.21). The resource allocation tokens are sent back to the egress node via the Reservation Start Ans message and forwarded in the RSVP [PATH] message to the transit domains(3). The ingress border node in each transit domain is responsible for the AA of the incoming RSVP request as well as resolving the path key to obtain the actual path inside the domain. The ingress border node contacts the local AAA system via the RSVP (R-TKN) Auth Req(4) which verifies the resource allocation token. If successful, it extracts the path keys and accounting parameters, and sends them to the requesting border node using the RSVP (TKN) Auth Ans message. The border node then queries the local PCE for the path using the extracted path keys (5) and uses the received path hop information for RSVP signaling inside the domain. This process is repeated in each transit domain until the destination. The standard RSVP procedure for resource reservation is then performed using the RSVP RESV signaling (9). After the RSVP signaling is completed, the source node uses a RSVP notification message to inform the local egress border node in the source domain of the successful completion of the reservation (10). The border node then notifies the local AAA server with a Reservation Finish message (11). Notice that only border nodes communicate with the AAA system in order to avoid impacting every router in the network and to reduce the management overhead for system administrators.

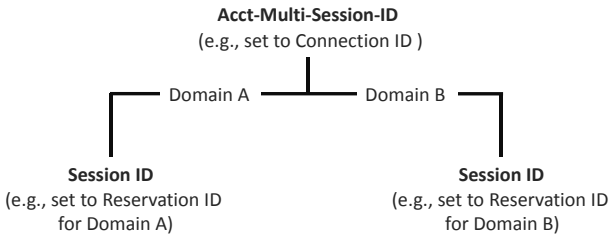


Figure 4.22: Correlating accounting records.

4.7.2.3 AA for Path Activation

Upon the successful completion of the path setup and when the source domain determines that the path needs be activated, the egress border node in the source domain contacts the AAA system to enable data forwarding in all transit domains using the *Path Activation Request* message (1). The AAA system instructs all transit domains' AAA systems to enable data forwarding using the *Open Gate* message which contains information that uniquely describe the connection such as the Connection ID, the Domain ID, and the Reservation ID which is used to reserve the path (see Fig. 4.21). The transit domains' AAA systems consult their policies, add any vendor specific attributes, and forward the request to their border nodes (steps 2-4). Once the source domain receives confirmations from all transit domains, it informs the requesting egress node that data forwarding is enabled for the path using the *Path Activation Completed(7)* message.

4.7.2.4 Accounting

Accounting can start immediately after the reservation operation is completed or after path activation. Each accounting session pertaining to each transit domain has a unique session identifier (Session ID) [45]. However, since different accounting sessions in different domains correspond to the same connection, we utilize the standard (Acct-Multi-Session ID) as a correlation identifier. This identifier is provided by the source domain in the resource allocation token during RSVP [PATH] signaling and is included along with the Session ID in all accounting messages. This idea is inspired by the concept of accounting record correlation that is widely used in cellular networks to correlate accounting records pertaining to multiple flows (voice, video, data) within the same session as well as records from different gateways if the session traverses through their respective service areas. For discussion purposes, let us first assume that the accounting process starts immediately after path activation. In this case, all metering border nodes within all domains send ACR Start messages to their local AAA systems to indicate

the beginning of the accounting session. The local AAA systems in all transit domains proxy accounting requests to the source domain (steps 1-6). After an accounting interim interval passes, accounting interim messages ACR (Interim) are sent to the local AAA systems which forward them to the source domain (step 7). Accounting interims are sent periodically during the connection lifetime to report the cumulative usage of the connection until it terminates. Once the connection terminates, ACR (Stop) messages are sent from all domains to report the total usage for the connection (step 8). Finally, in the case the accounting process starts immediately after path reservation, ACR(Start) messages are sent after the reservation phase and ACR(Interim) or ACR(Event) messages [45] are used to indicate path activation. Afterwards, the accounting signaling shown in Fig.4.20 is followed.

4.7.3 Security Discussion

The proposed AAA scheme for the PCE framework is expected to offer features of request authentication, authorization, integrity, non-repudiation, inter-domain path activation, topology hiding, and secure accounting as follows,

- Authentication: Digitally signed tokens are used to authenticate PCE and RSVP requests pertaining to a connection.
- Authorization: The requested QoS levels are allowed by the AAA system only if covered by the service level specification.
- Integrity: Digitally signed hashes protect the message contents (i.e., tokens) from modification.
- Non-repudiation: Digital signatures prevent repudiating actions relevant to path computation and reservation.
- Inter-domain path activation: Gate control mechanisms allow operators to control data forwarding over a reserved connection. For instance, it is possible to temporarily suspend data forwarding on some paths in response to denial of service attacks.
- Topology hiding: This is achieved by exchanging path keys between domains rather than the paths themselves and hence network topologies are kept private. Unused path keys are removed after the expiration of a negotiated timeout period.
- Secure multi-domain accounting: Accounting information is encrypted per Diameter standard and are directly sent from transit domains towards the source domain. We do not recommend proxy chain configurations in which transit domains forward accounting information for each other towards the source domain as it is possible for a compromised transit domain to modify accounting records

of the others. In the case that such chains are required, end-to-end privacy procedures should be applied according to the recommendations from the Diameter standard [45].

Now that we have described the security features of our framework, we discuss potential attacks and describe how our AAA framework mitigates them,

- Replay attacks: Replay attacks are not possible as we use sequence numbers for path computation and reservation. Since gate control and accounting messages are transmitted using Diameter, they are resistent to such attacks.
- Connection tampering: Transit domains cannot compromise the connection's security, such as when a compromised domain wishes to swap computed path keys for connections that were not reserved yet in order to spark billing disputes. This is counter-measured in our scheme because the path allocation tokens for a computed path are signed by the source and include unique connection identifiers. Hence, it is not possible for a compromised transit domain to swap allocation tokens for two valid connections. Accounting records also indicate the delivered QoS and path from all transit domains and hence can reveal any tampering.
- Connection theft: A connection theft attack is posed by a compromised transit domain which forwards RSVP [PATH] signaling from the source to the other domains while sending back an error indication to the source domain during the path reservation phase. The other domains are hence fooled into billing the source domain for the resource usage. This attack is mitigated by gate control mechanisms as the paths to each transit domain are only activated by the source domain using Diameter signaling which can not be recreated by the malicious transit domain. In addition, accounting records from all transit domains are sent to the source domain and hence can clearly expose such attacks.
- Denial of service attacks: We rely on standard intrusion prevention systems to minimize risks of denial of service.

4.7.4 Scalability Analysis

In this section, we study the scalability of the proposed AAA framework in terms of network and protocol parameters as well as service duration statistics. We provide a semi-analytic model for the mean AAA signaling rate (ξ) for the signaling flow in Fig. 4.20. As in the fixed model discussed in Section 3.4, the mean AAA signaling rate is generally described as the sum of the authentication and authorization (AA) ζ_{auth}, and the accounting signaling ζ_{acct} from all domains as,

$$E[\xi] = E[\xi_{AA}] + E[\xi_{Acct}] \tag{4.19}$$

The AAA signaling at any domain corresponds to the inter-domain connections generated from within the domain under consideration as well as connections generated in other domains and transit or terminate at the domain under consideration. We simply refer to the first as source connections and the latter as transit connections.

For source connections, let us assume that each domain generates connection requests at the rate of λ from C_i classes of service with proportions of $p_i\lambda$ for each class. Thus, given the mean hop count h_d per connection for all possible connections in the topology, the mean connection arrivals rate from service class i is $h_d\lambda p_i$. Notice that assuming the knowledge of the mean number of domain hops per connection, which is a function of the network topology, makes our model semi-analytic.

Now let us calculate the mean number of transit connections, t_d, that passes in a domain i ($\bar{X}_T^{(i)}$) given the total number of domains, N_d. To do so, let us denote the probability $a_{i,j} = \Pr\left\{\bar{X}_T^{(i)} \mid j \text{ is source}\right\}$. $a_{i,j}$ is topology dependent and can be obtained by simulations or measurements by counting all possible transit connections passing through a domain i given that the source domain is j. We denote the result as $k^*_{i,j}$. Assuming shortest path routing which is common in practice, the total possible number of paths from source j to all other domains is $N_d - 1$. Then, $a_{i,j}$ is given as $a_{i,j} = \frac{k^*_{i,j}}{N_d-1}$. Thus, for all possible sources and assuming uniform load, the total fraction of transit connections in domain i, denoted as a_i, is given as $a_i = \sum_{j=1,j\neq i}^{j=N_d} \Pr\left\{\bar{X}_T^{(i)} \mid j \text{ is source}\right\} \Pr\{j \text{ is source}\} = \frac{1}{N_d-1}\sum_{j=1,j\neq i}^{j=N_d} a_{i,j}$. Since we have (N_d-1) transit domains, the mean number of transit connections, t_d, in any domain i, is given as $t_d = (N_d-1)a_i = \sum_{j=1,j\neq i}^{j=N_d} k^*_{i,j}$.

Let N_v denote the mean number of switches that authenticate with the AAA system during RSVP signaling and N_g denote the mean number of switches that require gate control. The AA signaling is the sum of the AA exchanges[6] in path computation, setup, and activation (see Fig. 4.20) from source and transit connections as $E[\xi_{\text{AA}}] = E[\xi_{\text{AA}}^{(\text{src})}] + E[\xi_{\text{AA}}^{(\text{tran})}]$, as,

$$E[\xi_{\text{AA}}] = \sum_{\forall C_i} \lambda p_i \left[2 + 2N_v + N_g h_d + t_d\,(2 + N_v + N_g)\right] \tag{4.20}$$

To evaluate the signaling due to connection accounting, let us first denote the number of edge switches that meter the connections in a domain as N_m (e.g., $N_m = 2$ for transit domain A and $N_m = 1$ in the destination domain B in Fig. 4.20). Let us also denote the mean connection duration for service class C_i as $E[S_i]$ and the corresponding accounting interim interval as $\Delta_T^{(i)}$. If the complementary cumulative distribution of the connection duration is denoted as $\bar{F}_{S_i}(x)$, then using the generic formulation in (3.15) in Section 3.4,

[6] In our analysis, we consider a request and a response pair as an exchange similar to Chapter 3

the accounting signaling rate $E[\xi_{Acct}]$, is approximately[7] given as the sum of accounting signaling for source and transit connections which include accounting starts, interims, and stops as,

$$E[\xi_{Acct}] = \sum_{\forall C_i} N_m \lambda p_i (h_d + 1 + t_d) \left[2 + E\left[\frac{S_i}{\Delta_T^{(i)}} \right] \right]$$
$$= \sum_{\forall C_i} N_m \lambda p_i (h_d + 1 + t_d) \left[2 + \sum_{j=1}^{\infty} \bar{F}_{S_i}\left(j\Delta_T^{(i)} \right) \right] \quad (4.21)$$

For an exponential connection duration, distributed as $\bar{F}_{S_i}(x) = e^{\frac{-1}{E[S_i]}}$ in (4.21), the accounting rate is given in closed form as,

$$E[\xi_{Acct}] = \sum_{\forall C_i} N_m \lambda p_i (h_d + 1 + t_d) \left[2 + \left(e^{\frac{\Delta_T^{(i)}}{E[S_i]}} - 1 \right)^{-1} \right] \quad (4.22)$$

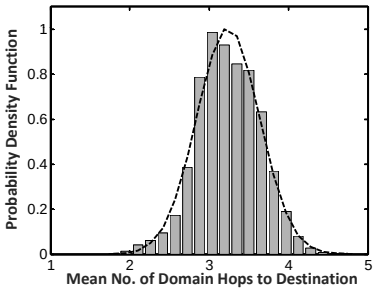
From (4.22), the accounting signaling load is proportional to the product of the number of edge switches that produce accounting (N_m), and the sum of mean domain hops per connection and transit connections $(h_d + t_d)$. It is also inversely proportional to the accounting interim interval setting Δ_T which determines the frequency of accounting updates per connection. Using a straightforward Taylor expansion of the $\left(e^{\frac{\Delta_T^{(i)}}{E[S_i]}} - 1 \right)^{-1}$ term in (4.22), it is easy to show that the accounting signaling complexity is $O((h_d + t_d) N_m \frac{S_i}{\Delta_T})$.

4.7.5 Simulations and Numerical Results

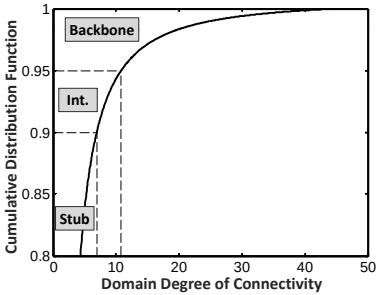
In this section, we study the scalability of our proposed AAA framework in terms of network and protocol parameters as well as service duration statistics. We use an event-driven simulator to measure the AAA signaling rate while varying the total number of domains in the system and the accounting interim interval settings. We generate 10,000 topologies using the BA algorithm in BRITE which follows the power law model [162][8]. For the generated topologies the average hop count is around 3 and ranges from

[7] This is because the destination domain and some domains may not have N_m switches which generate signaling.

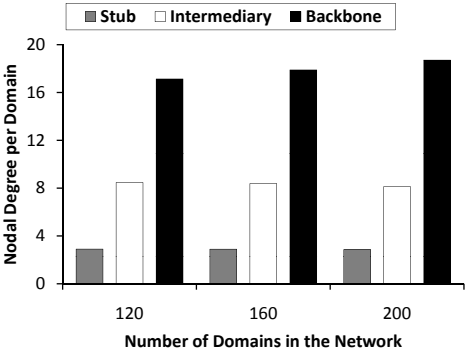
[8] In this model each new added domain has m links added to the existing domains with likelihoods proportional to their degree of connectivity. The number of links is given as $L = 2(N_d - 2) + 1 = 2N_d - 3$. The average degree of connectivity per domain is then $2L/N_d = 4 - 6/N_d$.



(a) Average hop count



(b) Avg. nodal degree for the 3 domain types



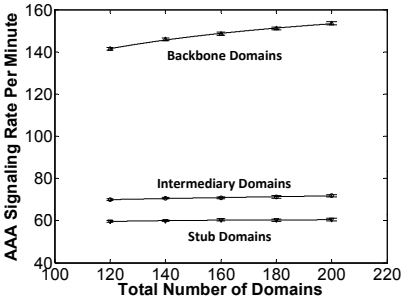
(c) Average nodal degree as function of domain type and network size

Figure 4.23: The generated topology statistics (BA topology generation model ($m = 2$)[162]).

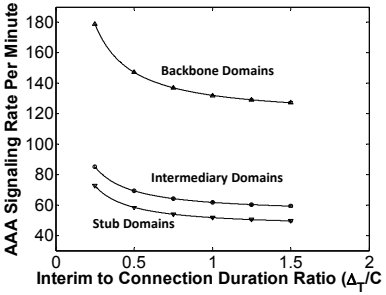
2 to 5 as shown in Fig.4.23(a). For each topology, we simulate AAA signaling due for a mixture of three cases which represent various services with long, medium, and short connection durations (e.g., enterprise, scientific applications, and video conferencing). The connection requests arrive according to a Poissonian process and the connection durations are exponentially distributed. Due to the power law model, the topologies follow a hierarchical nature[9] in terms of the number of peerings per domain (i.e., nodal degree). Hence, based on the distribution of the nodal degree, we categorize domains in the network into three types: (i) backbone [top 5%] (ii) intermediary [next 5%] (iii) stub [the rest, i.e., 90th percentile] as shown in Fig. 4.23(b). The average nodal degree

[9]In a hierarchical topology, few backbone domains transit most of the connections initiated by a majority of stub domains

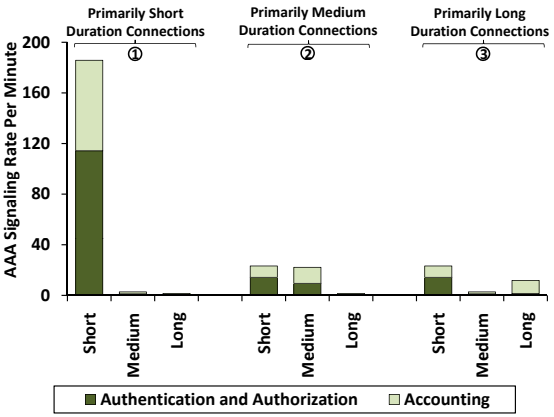
for each category for various network sizes is illustrated in Fig. 4.23(c).



(a) Load as a function of the number of domains



(b) Load as a func. of the interim interval



(c) AAA signaling load at backbone nodes as a function of different mixes of the short, medium, and long services. (1) 80%, 10%, 10% (2) 10%, 80%, 10% (3) 10%, 10%, 80%

Figure 4.24: Average AAA signaling load in Inter-Carrier Optical Networks [10,000 random topologies, 3 services with connection durations, C, of 2hrs (short), 1 day (medium), 1 week (long) with load shares per service as 60%, 30%, 10% resp (for scenarios (a), (b)) and accounting interim settings, Δ, in (a) and (c) are {1hr, 4hr, 4hr} resp. Total load from all services per domain = 150 Erlangs, 95% confidence intervals]

For the simulated topologies, the average number of domain hops is lower for backbone nodes as they have a higher nodal degree. The hop count is observed as 3.3, 2.8, and 2.4 for stub, intermediary, and backbone domains. From Fig.4.24(a), we see that the backbone domains experience higher AAA signaling due to the large number of transit

connections. The AAA load is seen to scale linearly as the total number of domains in the system increases. The observed load in Fig.4.24(a) is very light when one considers the fact that today's commercial AAA systems as in [21] can process hundreds of requests per second. Hence, we do not expect that the AAA system would be a performance bottleneck for our proposed optical application.

From Fig. 4.24(b), we observe that relatively low interim interval settings (i.e., below half of the average connection duration) result in a non-linear increase of the AAA signaling load. This aspect might not cause large AAA signaling load for the long lived connections with very low arrival rates, however this might not be the case for short lived connections with potentially significant arrival rates. To get further insight on the effect of the connection duration, we investigate the signaling load on the AAA system when a connection type dominates. From Fig. 4.24(c), we see that in all investigated traffic mixes, short lived connections result in significant AAA signaling load due to their higher arrival rates. We also notice that for dominant medium and long connections (cases 2, 3), the accounting traffic is more significant than the AA signaling since the interim setting is lower than half the connection duration. However, the overall AAA signaling load of such connections is low due their relatively low arrival rates.

To sum up, although a more in-depth analysis requires a wider security and signaling performance evaluation, our preliminary results in Fig. 4.24 are promising as they indicate that the AAA signaling load of our method is quite low even in relatively large topologies.

4.7.6 Open Issues

The following list illustrates some open challenges for future research relevant to this AAA application.

- *Redundancy*: The proposed signaling framework needs to be extended to authorize and configure failover paths.
- *Supporting online charging capabilities*: Online charging systems allow tighter control of resource usage as they allow almost realtime monitoring of charges relevant to resource usage.
- *Integration with connection monitoring systems*: Deep packet/frame inspection and tapping techniques can be used to monitor the quality of connections. Integrating such capabilities with the AAA framework allows better enforcement of Service Level Agreements (SLAs) and can be very useful from auditing perspectives.
- *Integration with network management systems*: In our signaling framework, we did not get into details of how and when RSVP or gate control signaling is trig-

gered. Integration with the network management systems can be a potential solution to provide this necessary information.

- *Integration with standards*: Our framework can be combined with evolving standards within the IETF for QoS authorization [163] and key management for routing and transport protocols (KMART) [164].

4.8 Conclusions

In this chapter, we developed novel mechanisms to optimize the performance of AAA signaling in terms of service authorization delay and accounting reliability in multi-service mobile network deployments. We also demonstrated the feasibility of AAA frameworks beyond traditional mobile networks in two novel areas including billing for cellular backhaul over wireless mesh networks and inter-operator layer 2 optical communications.

In Section 4.3, we studied the mitigation of QoS authorization signaling delay for services in emerging multi-service mobile architectures. Our motivation is that QoS signaling can result in undesirably variable and prolonged signaling delays upon handoffs between AGWs. This is because the policy system residing in the service tier may need to contact one or more application servers for QoS authorization which leads to degrading the quality of real time services. To address this important issue, we proposed a proactive signaling mechanism in the application layer that conveys authorization delay constraints from the service tier to the radio layer and thus mitigates the effects of variable signaling delay. The AAA system acts as a bridging network component which facilitates the communication of the delay constraints and proactive triggers between both tiers. The proposed mechanism is feasible as it uses the already established mechanisms of authentication and authorization signaling over standardized interfaces and protocols. Future research avenues for this work include integrating it with emerging standards such as the IEEE 802.21 framework, incorporating prepaid systems, and supporting policy interworking between heterogeneous technologies.

In Section 4.4, we showed that it is difficult to optimally set the accounting interim intervals in multi-service environments such that the incurred capital losses are minimal if the serving NAS fails. This is because although short accounting interim intervals minimize the potential loss, they are likely to result in undesirably high AAA signaling load especially when multiple services are used. Current accounting standards offer no quantitative measures for selecting proper reporting intervals and leave them open to the designers' choices. We showed that this problem is non-trivial as it primarily involves considering cost and statistical properties of multitudes of services with different mobility profiles as a multi-commodity optimization problem. To this end, we proposed a dynamic optimization mechanism to optimally trade off accounting reliability and AAA signaling load in multi-service deployments based on two policies: one which limits the

potential loss to a given value (a.k.a., CLP) and another which minimizes the potential loss without overloading the system (a.k.a., APWC). The proposed mechanism largely reduces the potential loss in the event of Network Access Server (NAS) failure without excessively generating unnecessary usage reports. It is based on IETF AAA standards such as RADIUS and Diameter and does not require changes to the network access servers in the network nor to the standards and changes are only limited to the AAA systems in the network. Future research avenues for this work include incorporating methods of estimation for the signaling costs of authentication and accounting messages, investigating the impact of load balancing between AAA systems, and extending the mechanism to optimize the operation of both the AAA system and business support system components in converged billing deployments.

In Section 4.5, we proposed the first accounting framework for cellular backhaul applications over wireless mesh networks and analyzed its scalability. Our proposal is motivated by the major challenge posed by today's backhaul networks, which connect base stations to the radio network controllers in cellular networks, due to the expensive maintenance and operations in response to traffic growth dynamics. Wireless Mesh Networks (WMN) were recently proposed to solve this issue due to their flexibility and low cost; however they raise new challenges including the maintenance of timing and synchronization, security, and billing. Relevant to our scope, we proposed an architecture where WMN operators offer backhaul services to cellular operators. In this regard, we designed an accounting mechanism in conjunction with a dynamic bandwidth management algorithm. The operation of the latter results in adding or releasing backhaul bandwidth chunks which generates accounting updates in the network. We considered two possible deployment cases: one in which only base stations are covered by the WMN and the Radio Network Controllers (RNC) are best reachable through the the WMN gateway node, and another in which both base stations and RNCs are reachable within the WMN. We also analyzed the billing signaling traffic resulting from bandwidth reservations based on a simple Markov process and evaluated a relatively unstable reservation mechanism. Based on our analysis, we established upper bounds on the accounting signaling performance for wireless mesh backhauls. We found that even a poor reservation scheme can be accommodated by current commercial hardware. Future research avenues for this work include designing lightweight protocol message formats to suite the WMN deployments, designing comprehensive network management solutions, optimizing the number and locations of the proposed intermediary mesh access points (IMAP), and performing a deeper investigation of data traffic variability on the AAA signaling load.

Finally, in Section 4.7, we proposed the incorporation of AAA signaling for layer 2 inter-carrier optical communications. Our proposal is motivated by the success of the Path Computation Element (PCE) framework which is used in multi-operator environments. We showed that this scenario is particularly interesting as it can be conceptually viewed as a generalization for the roaming scenarios in wireless networks as the data path may traverse more than two networks. As such this generalized case necessitates that the participating networks authorize path provisioning (i.e., data path computation

and setup) and also implement network access server (NAS) metering functionalities by some of their border switches. The proposed mechanism addresses such challenges and was specifically designed to facilitate secure exchange of path computation signaling among domains, associate path setup with the paths computed by the PCE, while enabling sharing of accounting information between carriers. The analysis showed that our signaling is light weight and may be integrated within the PCE platform, which demonstrates potential for commercial deployments. Future research avenues for this work include integration with connection monitoring systems as well as network management systems, securing the configuration of redundant paths, and supporting online charging paradigms.

To sum up, in this chapter, we have proposed AAA optimization mechanisms and future applications in the areas of cellular backhaul and optical networks. Our results demonstrated the promising research outcome in these areas towards standardization and commercialization.

Chapter 5 Results and Discussions

In this Chapter, we demonstrate the applicability of the AAA system planning models which we developed in Chapter 3 for centralized and distributed AAA system deployments. Afterwards, we turn our attention to performance optimization issues for authentication and accounting signaling which are discussed in Chapter 4. We first discuss the performance of the handoff QoS signaling delay minimization scheme and then elaborate on the performance of the accounting reliability optimization mechanism.

5.1 Simulation Model for AAA Signaling

In order to validate our AAA system planning models, we extended the call/session-based simulation model described in [165] to incorporate AAA protocol operation as well as AGW residence times. In our simulation model, we consider protocol events in addition to the standard three event types of new session arrivals, handoff, and termination. The self-explanatory Procedure 1 summarizes the steps needed to handle each event case. An interesting fact to note is the generation process of the residence time and its residual for AGWs. The generation of the residence time was carried out by simulating multiple cells per AGW with lognormal residence time for each cell, with users initiating movement on the border cells. The samples of the residence time until the user departs the AGW area were recorded into multiple files which are long enough to cover the number of samples needed by simulations.

The residual residence time for the AGW was obtained similarly by initiating movement inside the AGW region under consideration and keeping record of the gateway residence time. The other method we also used is to directly generate the residence time as a random variable and its residual using the procedure in [165]. For this method, a noteworthy implementation observation we found, is that the statistics generated by the residual of the lognormal distribution were not as stable as other distributions such as the residual gamma for instance. In our simulations, sessions leaving the network are not traced and offnet arrivals are assumed constant for comparison purposes.

The simulation logic is described in the following procedure.

Procedure 1 Extended Call/Session-Based Simulation Model for AAA Signaling

Input : p_a and p_d access success and session dropping probabilities resp., interim interval(Δ_T), reauthorization life time (Δ_M), and session arrival rate in the network.
Initialize: Generate *a session arrival*, set event type to *new session*, and add it to the event list
foreach *event in the event list* **do**
  switch *event type* **do**
    case *new session event*
      handleNewSessionEvent()
    case *protocol event*
      • Update statistics for the serving AGW and AAA systems
      • *handleProtocolEvent()*
    case *handoff event*
      • Session dropping: Generate a uniform variable x and compare it to p_d.
      • *handleHandoffEvent()*
      • Update statistics.
    case *session termination event*
      • *handleSessionTerminationEvent()*
    otherwise Indicate an error. This case reports any abnormal execution conditions.
  end
end

The *handleNewSessionEvent()* function is defined as follows,

Procedure 2 Handling New Session Events: *handleNewSessionEvent()*

- Admit session if resources per cell are available (we assume that this always works).
- Authenticate the call: send authentication request and generate a uniform random variable (R.V), x, if $x \leq p_a$ then accept.
- Schedule accounting request (start) one round trip after the current time (t_c) as ($t_c + RTT$) where RTT denotes the round trip between the AAA and the AGW.
- Generate random number for the session duration S, the starting *AGW* index, i, according to the users' distribution in the network.
- Set the current residence time t_r to the initial AGW residence time as $t_r = \tilde{R}_i$ where i is the index of the AGW (we assume different AGW residence time distributions).
- Schedule Protocol Updates: (a) schedule interim if $\min(S, t_r) > \Delta_T \Rightarrow$ schedule protocol event type interim. (b) Schedule reauthentications similarly to accounting interims but using the Authorization-Lifetime (Δ_M) instead of (Δ_T).
- If ($S > t_r$), schedule a handoff event; otherwise a session termination event.
- Update statistics for the serving AAA, as multiple AAAs can be used, and AGW (e.g., arrival rate, authentication accept rate, etc).

The *handleNewProtocolEvent()* function is defined as follows,

Procedure 3 Handling Protocol Events: *handleNewProtocolEvent()*

switch *protocol event type* **do**
  case *Interims*
    if $t_c + \min(S, t_r) > t_c + \Delta_T$ **then** schedule protocol event type interim.
  case *Re-authentications*
    if $t_c + \min(S, t_r) > t_c + \Delta_M$ **then** schedule protocol event type re-authentications.
  otherwise Handle any other messages here.
end

The *handleHandoffEvent()* function is defined as follows,

Procedure 4 Handling Handoff Events: *handleHandoffEvent()*

if $(x \leq p_d)$ **then**
  Invoke *session termination event* with status "dropping".
else

- Set the remaining session duration $S = S - t_r$, use the mobility model to determine the next AGW (AGW j). Generate an AGW residence time sample R_j and set the current residence time t_r to the generated sample $t_r = R_j$.
- Protocol updates: Send accounting stop message by the source AGW. Send authentication request and schedule accounting start by the target AGW after a round trip delay with its AAA system.
- Schedule Protocol Updates: (a) schedule interim if $\min(S, t_r) > \Delta_T \Rightarrow$ schedule protocol event type interim. (b) Schedule reauthentications.
- If $(S > t_r)$, schedule a handoff event; otherwise a session termination event

end

The *handleSessionTerminationEvent()* function is defined as follows,

Procedure 5 Handling Session Termination Events: *handleSessionTerminationEvent()*

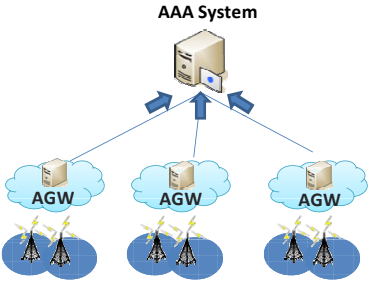
- Protocol updates: Send accounting stop message by the current AGW.
- Update statistics (indicate if dropped call or normal termination).

5.2 AAA System Planning: Centralized Deployments

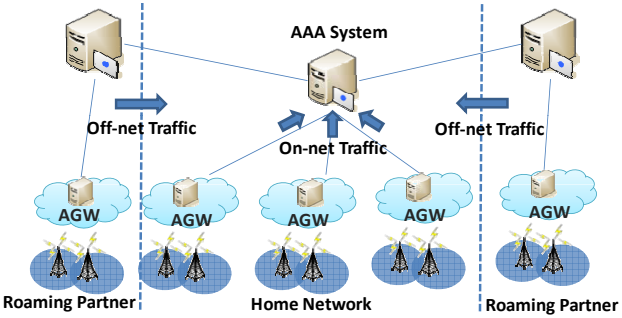
In this section, we start by studying the AAA signaling rate due to home users in mobile networks and compare it to the corresponding signaling rate in fixed networks. We assume a centralized AAA system serving all users in the network (see Fig. 5.1(a)) and that residence times are i.i.d in AGW areas. Afterwards, we study the signaling rate

behavior due to roaming users compared to home users and investigate the suitability of the basic model in Section 3.5 as an approximation for roaming scenarios. For roaming users, we assume two operators situated east and west as shown in Fig. 5.1(b).

Our simulations are based on the logic described in Procedure 1 in which we simulate macro cell sizes with users following fully random mobility between the cells. The number of cells and the parameters of the residence times are varied to reflect different AGW residence times. In our results, we study a range of mobility profiles which are characterized by the ratio of the mean session duration to the mean residence time.

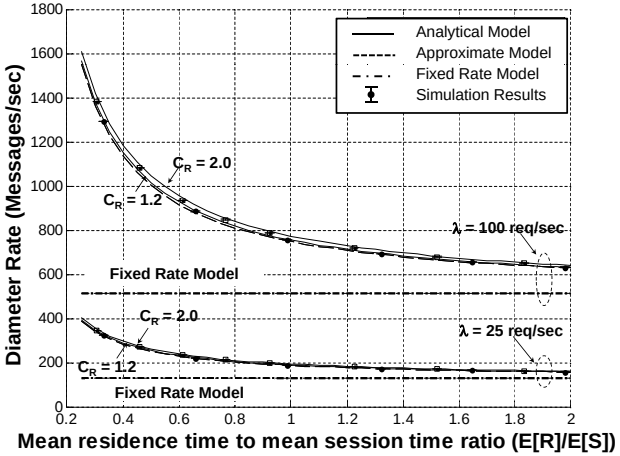


(a) Scenario A: Home Network

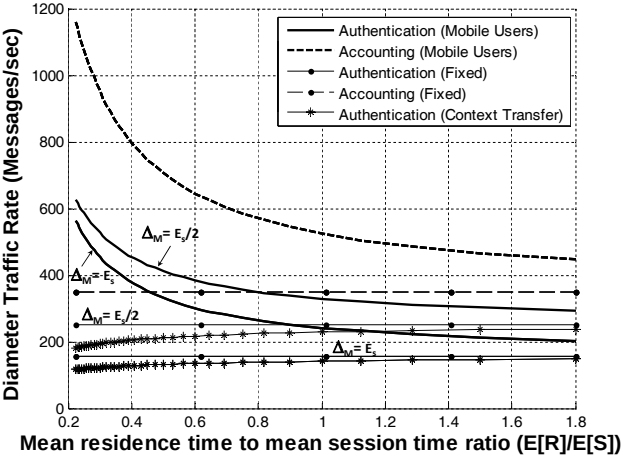


(b) Scenario B: Roaming Users

Figure 5.1: AAA system planning model (centralized AAA systems).



(a) Residence time effect on the mean signaling rate



(b) Residence time effect on the authentication and the accounting traffic

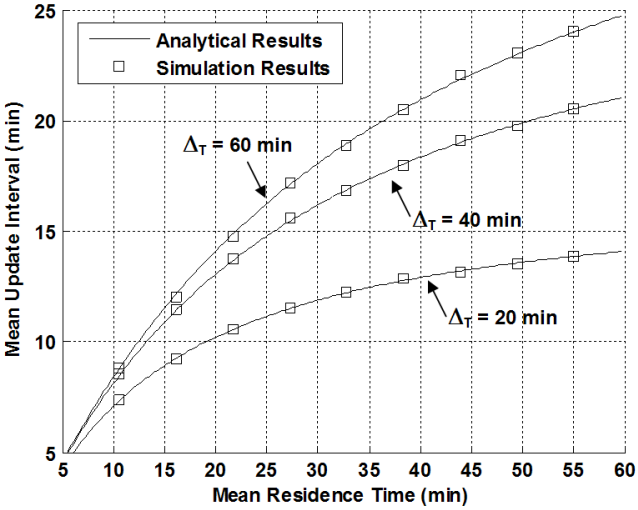
Figure 5.2: Mean AAA signaling load for home users (adapted from [102]) [Simulation parameter: In Fig.5.2(a), 5 AGWs, $E_S = \Delta_M = 40$ min, $\Delta_T = 20$ min, 5×5 cells per AGW with residence times varying from 2.5 - 15 mins per cell (lognormal coeff. of var. $\in \{2,3\}$). Mean batch method (30 batches, 10 hr simulation, 95% confidence) (error bars are within the marker sizes). In Fig.5.2(b), E_s = 40 min, C_R = 2, $\Delta_T = E_s/2$, $\Delta_M \in \{E_s/2, E_s\}$.].

5.2.1 The AAA Signaling Rate Due to Home Users

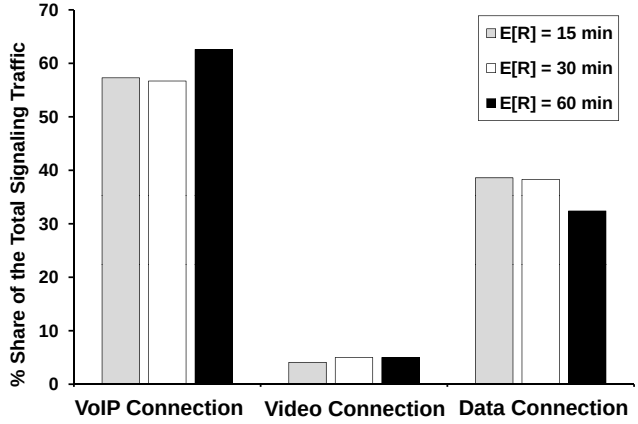
In Fig.5.2(a), we show the effect of the residence time on the mean AAA signaling rate. We compare simulation results with the *basic model*, *its approximation*, and *fixed network models* in (3.38), (3.39), and (3.11). Our goal is to obtain an understanding of the conditions that each model applies to. We see that the analytical results (lines) match simulation results (dots) within $< 2\%$ error. The approximate model also gives good estimates ($< 5\%$ error). This comes from the observation that changing the coefficient of variation ($C_R = k_r^{-0.5}$) for the residence time does not change the resulting signaling rate considerably, suggesting that exponential approximations for the residence time perform reasonably well. We also notice that as the residence time to session duration ratio increases, the signaling rate approaches the fixed rate model asymptotically. This is because when residence times are large, handoffs are unlikely and hence the fixed rate model applies. We observe that for low mobility (i.e., when the mean session is approximately half or less than the mean residence time, the fixed AAA signaling model can be used with an error less than 20%. However, for fast moving users or users located on the borders, with mean residence times less than half the mean session duration, the basic model in (3.38) should be used instead.

In Fig.5.2(b), we show the effect of changing the mean residence time duration characterizing mobility on the authentication and the accounting traffic loads for two different values of the authorization lifetime. We see the traffic split and observe that the accounting traffic load is higher than the authentication traffic load for the used authentication scheme (i.e., the 3GPP2 CHAP based authentication in [44]). For practical authorization lifetime settings, the number of accounting messages is usually larger than authentication messages. Again, we see that using the fixed model results in a large estimation error when the ratio of the mean residence time to the mean session duration is low (i.e., high mobility). This is because for both authentication and accounting traffic the ratio $\frac{E[S]}{E[R]}$ is a scaling factor. For the case when the context is transferred between AGWs, if the reauthentications are triggered based on the session start time, the observed authentication rate matches that of a fixed network. This is because in this case, mobility does not trigger authentications nor change the timing of reauthentications. On the other hand, if reauthentications are triggered based on handoff instants rather than the session start time, the corresponding authentication rate approaches that of the fixed model. This is because when the E_r/E_s ratio is low, reauthentications are barely triggered. However when $E_r > E_s$, the number of re-authentications is limited by the session duration rather than the residence time. For context transfers and practical settings for the authorization lifetime (e.g., $\Delta_M = E_s$ and $\Delta_M = .5E_s$), we conclude that the model for fixed networks in (3.11) can be used to estimate the authentications rate even for relatively high mobility.

In Fig.5.3(a), we study the effect of the mobility on the time between accounting updates. This is an important metric as it directly relates to the risk of accounting losses for a given mobility profile at planning time. Clearly, the increased mobility (i.e., smaller



(a) Mobility effect on the mean update interval



(b) Service shares of the mean signaling rate

Figure 5.3: Mean AAA signaling load for home users (adapted from [102]) [Simulation parameter: In Fig.5.3(a), $E_S = \Delta_M = 40$ min, $k_r = 0.25$, 100,000 sessions. In Fig.5.3(b), $\Delta_T \in \{2.5, 10, 30\}$ and $\Delta_M \in \{5, 20, 60\}$ for VoIP, video, and data respectively].

residence times) results in a reduced accounting update interval and hence in increased reliability. This is due to the higher likelihood of accounting stop records which may occur more frequently than the interim updates. However, shorter update intervals may reflect on higher capacity requirements for gateways relying on such accounting updates as well as on higher AAA system capacity. Using the same reasoning, one observes that as Δ_T/E_r ratio decreases, the interim updates will dominate and hence the update interval approaches the interim interval Δ_T. This result unveils a challenge for the choice of the interim interval as the mobility statistics can vary during the day which can turn the use of the interim interval excessive or useless if not properly adapted to the observed mobility and session statistics. We address this issue in Section 5.5.

Results shown in Fig.5.3(b) illustrate the effect of session durations on the mean signaling rate, for an exemplary mix of services of 70% VoIP, 5% video, and 25% data with session durations of 5, 20, and 60 mins, respectively. The total signaling rate is the sum of the rates from all services, assuming that radio resource admission and allocation are always successful. The signaling rate due to each service is calculated based on (3.38) with the corresponding arrival rate $\lambda^{(i)}$ values set according to the service traffic proportion, and the interim interval and the authorization lifetime set to half and full session durations (i.e., $0.5E_s$ and E_s) respectively. As shown in Fig. 5.3(b), the signaling rate shares are not necessarily proportional to the service arrival rates. In other words, we neither have AAA rate shares of 70% for VoIP (i.e., 57-63% instead) nor 25% for data (i.e., 32-38% instead) leading to a margin of 15% from the service shares. We notice that higher residence times result in an increase in the share of the short duration services (i.e., VoIP) while reducing the shares for longer session durations (i.e., data) due to the lower number of handoffs. We hence conclude that for mobile users even if one sets their interim and reauthorization lifetime intervals proportionally to the service session duration, this is does not mean that the AAA signaling rates will follow the arrival rate proportions of the services.

Finally, we compare the AAA signaling rate obtained by the basic model under exponential session duration assumptions with the results obtained by simulating lognormally distributed session times (i.e., non-exponential) with relatively large coefficient of variation ($C_s = 2$). Table 5.1 shows that the error between the analytical model and the simulations is less than 13%. This error is due to the estimation of the length of the holding times. We argue that such error due to the exponential distribution can be considered within the practical 20% design margin and does not necessarily result in excessive over provisioning of the system. Therefore, the exponential model offers reasonably tolerable accuracy, even for generic sessions with high variance ($C_s = 2$).

5.2.2 The Impact of Roaming in Centralized AAA Systems

We now study the behavior of foreign (i.e., MVNO/roaming) users in mobile networks using our generalized model in Section 3.6 in equation (3.73). The goal here is to

Table 5.1: A comparison between the basic AAA model in (3.38) and simulations with lognormally distributed session times with coefficient of variation of 2 (adapted from [102]) [E_r = 18.4 min, C_R = 2, $\Delta M = E_s$, λ = 100 req/sec, mean batch method, 30 batches, 95% confidence].

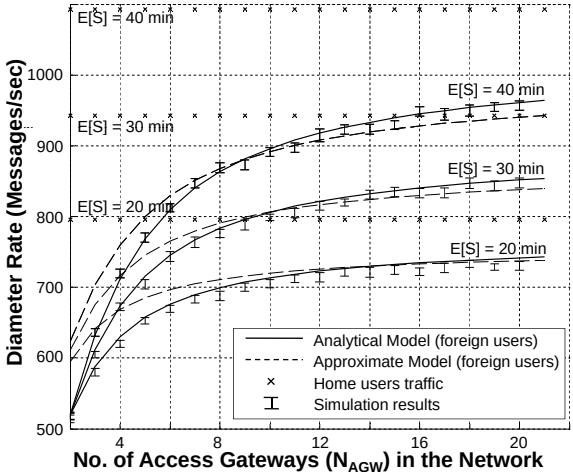
Interim Interval (Δ_T)	**E[S] = 40 min**		**E[S] = 30 min**		**E[S] = 20 min**		**E[S] = 5 min**	
	Ana.	Error	Ana.	Error	Ana.	Error	Ana.	Error
$\mathbf{E_s/4}$	1278	7.42%	1132	6.78%	987	6.95%	777	6.90%
$\mathbf{E_s/2}$	1093	8.60%	943	6.41%	796	9.01%	580	11.61%
$\mathbf{E_s}$	1013	7.41%	860	6.57%	708	6.73%	486	12.27%

form a basic understanding of how the signaling rate varies with the network size (i.e., number of AGWs) and the users' distribution in the network. The topology is based on Fig.5.1(b). For comparison purposes with the basic model, we fix the on-net and off-net traffic[1] proportions to 80% and 20% respectively. This means that the on-net and off-net traffic rates do not change as the number of AGWs is increased. Our goal here is to illustrate the estimation error incurred when planning AAA systems for roaming users even under high mobility assumptions ($Es/Er = 2.17$). We simulate at relatively high arrival rates (100 req/sec) to allow comparison with the results in Fig.5.2. For the signaling rates due to home users (predicted by the basic model in (3.38)), and the fixed network model in (3.11), the arrival rates are set such that $\lambda = \lambda_\Omega + \lambda_\Phi$ = 100 req/sec.

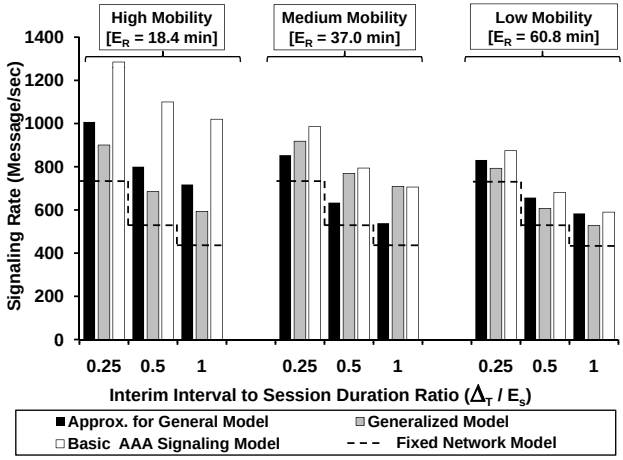
Fig.5.4(a) shows the signaling rate due to the foreign traffic as a non-linear monotonically increasing function of the network size, i.e., the number of AGWs. The trend is followed even though the on-net and off-net arrival rates are fixed. We see that simulation results match the analytical model well within the 95% confidence and that the approximate model, which assumes exponential AGW holding times, provides good results for networks with a large number of gateways (i.e., > 6). We also notice that the signaling rate predicted by the basic AAA model is unaffected by the number of AGWs in the network. This is because the basic AAA model assumes the reception of the AAA signaling for the *whole session*. This is not the case for foreign users who may leave the visited network (i.e., the network under consideration). Thus, as the number of AGWs in the visited network is increased, the tendency for foreign users to leave the network decreases, and hence the AAA signaling rate approaches the rate predicted by the home AAA model. This is confirmed by the observation that shorter sessions approach such limits faster as they allow a smaller number of handoffs and hence more likely to terminate within the network.

In Fig.5.4(b), we study the impact of interim interval and mobility on the signaling rate observed by the visited network's AAA system. In our analysis, the interim interval is normalized to the session duration ($\frac{\Delta_T}{E_s}$) to obtain general trends. We first see that low

[1]Recall that on-net traffic refers to signaling due to roaming sessions starting from within the roaming operator's network while off-net traffic refers to signaling due to roaming sessions starting outside the network under consideration and then handing off into it.



(a) AAA signaling load as a function of the number of access gateways



(b) AAA signaling load as a function of the normalized interim interval Δ_T

Figure 5.4: Mean AAA signaling load for roaming users [In Fig.5.4(a), E_r = 18.4 min, C_R = 2, $\Delta_M = E_s$, $\Delta_T = 0.5E_s$, $\lambda_\Omega = 80, \lambda_\Phi = 20$ req/sec. In Fig.5.4(b), E_s = 40 min, $N_{AGW} = 10$. Mean batch method (30 batches, 95% confidence intervals), 10 hr simulations in Fig.5.4(a)- 5.4(b)].

interim-intervals result in a large signaling rate due to the frequent accounting interim updates and that higher mobility results in higher signaling rate due to the more frequent AGW handoffs, as expected. Similar to Fig.5.4(a), we also observe that the generalized model's results are closer to the basic model's predictions in medium and low mobility scenarios. It is also close to the fixed model in low mobility scenarios. Moreover, the error margin is lower for low interim settings (e.g., see the high mobility case with $\frac{\Delta_T}{E_s} = 0.25$ compared to $\frac{\Delta_T}{E_s} = 1.0$). From Figs.5.4(a)-5.4(b), we conclude that the basic model serves as an upper bound on the signaling rate from roaming users while the fixed model is a lower bound. Under uniform user distribution assumption, we also conclude that for small operators supporting services with long session durations and interim settings that are not too small ($\frac{\Delta_T}{E_s} \geq \frac{1}{2}$), planning AAA systems for roaming users similar to home users can result in large over provisioning. On the other hand, such effects do not pose over-provisioning issues for large operators.

Let us now see how the foreign users' distribution among AGWs plays a role in determining the AAA signaling rate. This scenario is more relevant to MVNOs as their users may be highly concentrated in one region relevant to their customer base. A uniform distribution is more likely in the case of traditional roaming partners because high users' concentrations would rather be covered by their networks (within the same country) to maximize profit. In our study, we focus on on-net traffic as off-net traffic is naturally concentrated at the network borders. We simulate at 8 sessions/sec for on-net traffic. Table 5.2 lists the signaling rates for various initial distributions and different interim intervals. We observe that the minimum signaling rate occurs when the users are totally edge concentrated (as the network departure probability is maximized) while the maximum signaling rate occurs when the users are completely concentrated at the central AGW (as the departure likelihood is minimized). Hence, the signaling rate due to uniformly distributed users lies between the edge and the center-concentrated extreme cases as expected. In all cases, the basic model is the upper bound of the signaling load regardless of the users' concentration. In summary, the more concentrated the users are to the center, the less likely that they leave the network, and hence the closer their signaling rate to the predictions of the basic model.

Table 5.2: The effect of the initial distribution of onnet traffic on the resulting AAA signaling rate for roaming users [$N_{AGW} = 5$, E[S] = 40 min, E[R] = 18.4 min, $C_R = 2$, $\Delta M = E_s$, $\lambda_\Omega = 8, \lambda_\Phi = 2$ req/sec, mean batch method (95% confidence intervals), 10 hr simulations].

Interim Interval	**Initial On-net Distributions (F_Ω)**					
	Basic Model	**Edge [1 0 ... 0]**	**Uniform [.2 ... .2]**	**[.1 .2 .4 .2 .1]**	**[0 .2 .6 .2 0]**	**Centered [0 0 1 0 0]**
$\mathbf{E_s/4}$	128.5	78.7	90.6	95.2	99.9	101.9
$\mathbf{E_s/2}$	110.0	66.2	76.9	81.0	85.2	87.0
$\mathbf{E_s}$	102.0	60.8	70.9	74.9	78.8	80.5

Finally, in Table 5.3, we study the quality of the general model's signaling estimates in (3.73) for a range of the number of AGWs (i.e., $N_{AGW} \in \{2, 20\}$). We compare

the general model's results to simulation results for lognormally distributed sessions (coefficient of variation $C_s = 2$). As shown in Table 5.3, the maximum error is below 14% and occurs at short interim intervals for large sessions. Even for relatively large mean session durations (40 min), the error of using the exponential distribution does not result in excessive over provisioning as in today's systems and can be considered within the practical design margin of 20%. Therefore, our model offers tolerable practical accuracy even for generic sessions with relatively high variance ($C_s = 2$).

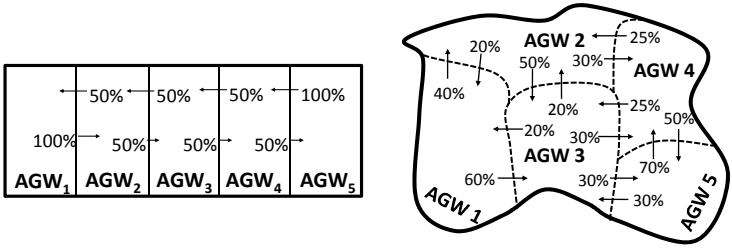
Table 5.3: A comparison between the generalized AAA model and simulations with lognormally distributed session times with coefficient of variation of 2 [$N_{AGW} \in \{2, 20\}$, E_r = 18.4 min, C_R = 2, $\Delta_M = E_s$, $\lambda_\Omega = 80, \lambda_\Phi = 20$ requests/sec, roaming w/o services (RWO) configuration, simulation's mean batch method is used (30 batches, 95% confidence intervals), 10 hr long simulations].

Interim Interval	**E[S] = 40 min**		**E[S] = 5 min**	
	Ana.	Error	Ana.	Error
$\mathbf{E_s/4}$	1054.9-1520.1	9.1-13.7%	945.0-972.3	0.6-5.0%
$\mathbf{E_s/2}$	882.4-1300.9	8.0-11.6%	703.6-730.3	0.6-3.6%
$\mathbf{E_s}$	806.6-1206.2	6.2-9.7%	588.1-614.8	0.1-2.0%

5.3 AAA System Planning: Distributed Deployments

The goal of this section is to investigate the applicability of the generalized planning framework in Section 3.6 using multiple numerical results. To this end, we first show that our general model can be used to determine how the AAA signaling load is distributed in an exemplary network of five AGWs served by a centralized AAA system and then show how it can be used to determine the AAA signaling load in a distributed AAA system deployment serving a network consisting of 16 AGWs with different residence times and authentication protocols.

In our numerical results, we use our framework to characterize the AAA signaling load in mobile networks with different topologies and mobility patterns serving home and roaming users. Moreover, we show that our model can handle practically interesting scenarios in mobile networks incorporating different cellular technologies (e.g., LTE and WiFi or EVDO and WiMAX) where different authentication protocol deployments within the network are used and non-identical AGW residence times are expected. We emphasize that our framework allows analytical consideration of relatively complex yet realistic deployments of distributed AAA systems under general assumptions of residence times and authentication protocol choices. Such results can only be obtained currently through exhaustive simulations or lengthy laboratory load tests.



(a) Mobility Pattern A (b) Mobility Pattern B

Figure 5.5: Topology and mobility patterns in a network of 5 AGWs (Centralized AAA System).

5.3.1 The Signaling Load Distribution Among Access Gateways

Consider a network consisting of five AGWs and served by a centralized AAA system. Our task is to study the amount of signaling load that each AGW generates towards the AAA system for two arbitrary[2] mobility patterns: Pattern A and Pattern B shown in Fig. 5.5. Using our generalized AAA model in (3.73) in Section 3.6, we study three user profiles with very low, medium, and very high mobility with corresponding mean session to AGW residence time ratios (E_s/E_r) of 0.1, 1.0, and 5.0 respectively. Fig. 5.6(a) illustrates the AAA signaling load from each AGW. For low mobility users, the load generated by each AGW in the network matches the session initiation probabilities irrespective of the mobility pattern as expected and similar to the behavior in fixed networks. On the other hand, for medium and high mobility, the load distribution among AGWs is no longer uniform due to the handoff sessions. For instance, AGW1 receives more handoff sessions in mobility Pattern A than in Patten B and hence the signaling load pertaining to Pattern A is larger accordingly. The opposite is true for AGW5. Clearly, as the mobility increases, the signaling load generated by each AGW becomes more dependent on the mobility pattern and hence the difference in the AGW load distribution for different mobility patterns becomes more visible. However, does the consideration of the mobility pattern change the *total* AAA signaling load from all AGWs in the network ?

To answer this question, let us consider the case for the two mobility patterns for roaming and non-roaming users and under variable mobility conditions (i.e., AGW residence times). In the roaming case, we allow sessions to leave each AGW by a given percentage by equally reducing movement in all other directions. As shown in Fig. 5.5(b),

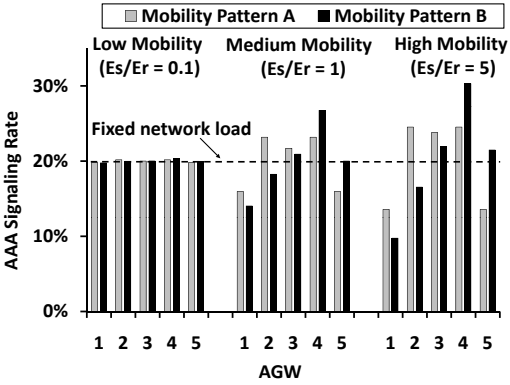
[2]In real networks, the mobility patterns can be practically deduced by observing the accounting records from all AGWs for all sessions and calculating the likelihood of movement between AGWs during the session lifetime. We assume uniform user distribution among AGWs.

for non-roaming scenarios the signaling load perceived by the AAA system is the same irrespective of the mobility pattern and equals that calculated by the basic model in (3.38). For roaming scenarios, when the two networks have the same roaming likelihoods from each AGW, the signaling load is the same regardless of the mobility pattern (see the 10% roaming case) while it differs otherwise (see the 5% roaming case). This is because when the roaming likelihoods are the same at each AGW then sessions roam outside the network with similar likelihoods irrespective of the mobility pattern inside the network.

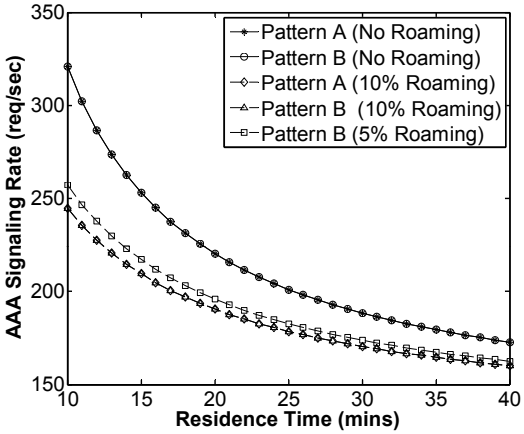
5.3.2 Impact on the AGW Holding Times

Let us now delve into more details of how mobility affects the AGW holding time durations. Recall from Section 3.5 that the AAA system perceives four AGW holding times: Full, Originating, Transit, and Terminating. These quantities are important as they directly determine the reauthentications and accounting interim signaling rates. AGW holding times are what an AAA system actually perceives from each AGW (see Table 2.1) and are distinguishable using the Begin-of-Session and the Session-Continue attribute value pairs [43, 44]. The knowledge of the statistics of each of these AGW holding times can guide the choice of the right accounting interim and authorization lifetime intervals. It can also facilitate designing effective post processing of accounting information where all records from a session are correlated into one standard (i.e., normalized) record. In Fig.5.7(a), we show the mean holding time duration for the four AGW holding times as function of mobility. In Fig.5.7(b), we show the corresponding occurrence likelihoods for each AGW holding time category. For low mobility, full sessions dominate and their mean duration is approximately equal to the mean session duration. For medium mobility ($Es/Er = 1$), the mean durations of the four session types as well as their occurrence likelihoods are very similar and are around half the session duration. For high mobility users, the holding time durations are lower as the users tend to make more handoffs during the session and hence the transit holding times dominate (see Fig.5.7(b)).

In order to see the impact of the mobility pattern on the holding times generated by each AGW, we apply the two mobility patterns A and B in Fig.5.5 and observe the percentage of each of the four types from each AGW. The session initiation probabilities are uniformly distributed in all AGW regions. For low mobility, full session holding times dominate and are uniformly distributed among all AGWs as shown in Fig.5.7(a). Notice that the distribution of the full sessions follow the session initiation likelihoods by definition and hence their dominance makes the load distribution among AGWs more similar to the session initiation likelihoods. In other words, the dominance of the full sessions is what makes the load distribution among AGWs uniform in our finding in Fig.5.7(a). For high mobility, we observe the dominance of the transit connections in both mobility patterns as shown in Fig. 5.7(b). However, a closer investigation of the holding times in each AGW area in Figs.5.8(a)-5.8(b) shows that the occurrence likeli-



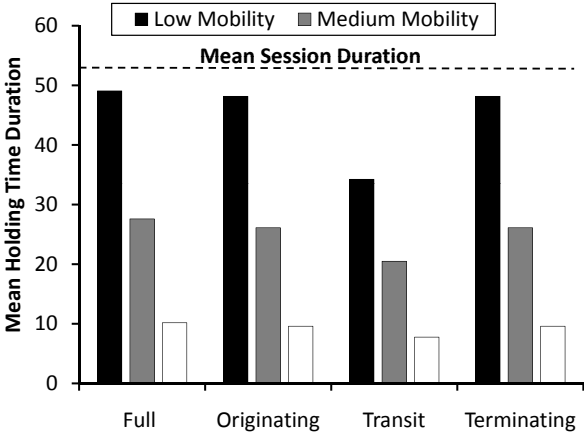
(a) Signaling load from each AGW



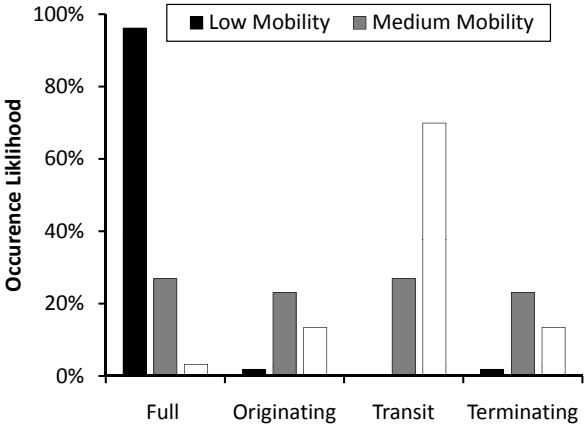
(b) Signaling load as function of mobility & roaming

Figure 5.6: AAA Signaling load and its distribution as function of mobility and roaming [$E_s = 50$ min, $C_R = 1.2$, $\Delta_T = E_s/2$, $\Delta_M = E_s$, and $\lambda = 5$ sessions/sec/AGW, %mobile users = 100%].

hoods of only transit and terminating AGW holding times are affected by the mobility pattern (i.e., see that the corresponding black and light gray bars are all of equal heights). This is because, as we stated in Section 3.6.2.6, the occurrence likelihoods of full and originating sessions only depend on the initial user distribution and not on the mobility pattern as the other types do.



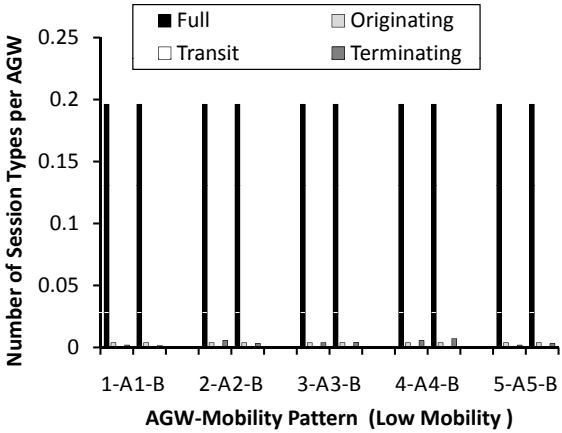
(a) Holding times duration as function of mobility



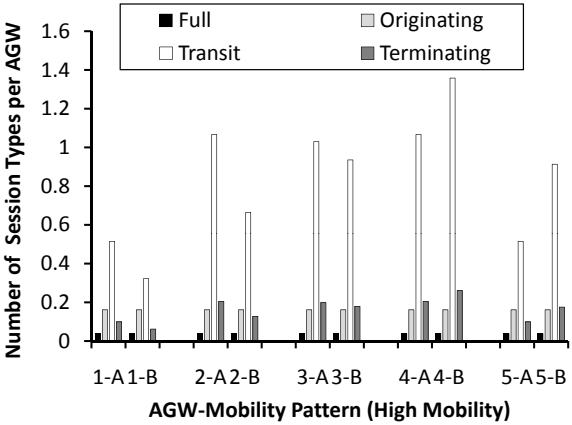
(b) Holding times occurrence likelihood

Figure 5.7: Holding times durations and their occurrence likelihoods.

To sum up, we showed that even though the AAA load is independent of the mobility patterns in some cases, the load distribution among AGWs as well as the likelihoods of occurrence for transit and terminating AGW holding times are always affected by the mobility pattern. We clearly demonstrated that only in relatively low mobility scenarios



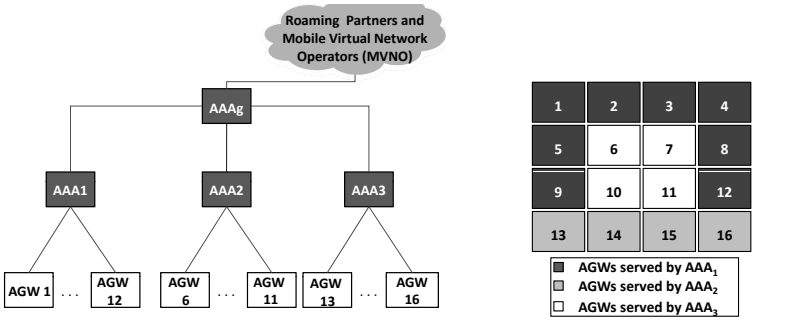
(a) Holding times occurrence distribution (low mobility)



(b) Holding times occurrence distribution (high mobility)

Figure 5.8: Holding times occurrence distributions among AGWs.

where full AGW holding times dominate can we say that the load distribution among AGWs matches the users' distribution in the network. Finally, we also demonstrated that the effect of the mobility pattern is nullified in roaming scenarios when the roaming likelihoods from each AGW are similar.



(a) AAA Architecture (b) 4×4 AGWs (AAA assignment)

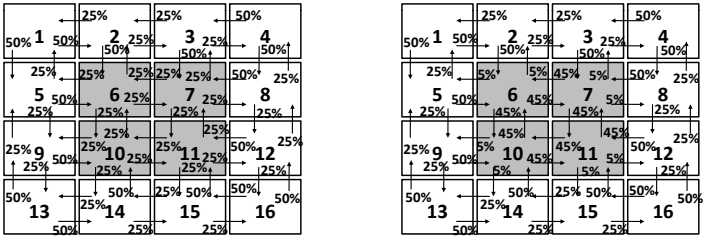
Figure 5.9: AAA architecture and assignment in a network of 16 AGWs.

5.3.3 The Signaling Load in Distributed AAA Deployments

In this section, we consider a 4×4 AGW configuration served by four AAA systems. AAAs 1-3 authenticate and collect accounting messages for home users and proxy the AAA signaling for roaming and MVNO users[3] to a gateway AAA system (AAAg). AAAg, in turn, forwards this traffic to the roaming partners and/or MVNO networks as shown in Fig.5.9(a). AAAg also handles AAA requests from home users - with respect to the network under consideration - while being served by roaming partners. To illustrate the applicability of our framework, we arbitrarily assign AGWs to the three AAAs as shown in Fig.5.9(b).

In our case study, we consider two mobility patterns: Pattern C where users move in all directions with equal likelihoods and Pattern D where users tend to go to the central AGWs as shown in Fig.5.10. We assume that only 30% of the users are mobile while the remaining 70% never leave their serving AGW. To investigate the impact of realistic cases where AGW residence times differ, the residence times of the center AGWs (6,7, 10, 11) are set to half the value of the rest. This in practice may reflect higher mobility in some AGW regions or even different cellular technologies resulting in a measurable difference in the residence time value. Before we proceed to the details of the signaling load at each AAA system, we first show how the difference in the residence time affects the load distribution among AGWs. For low mobility scenarios, the load distribution among all AGWs matches the user distribution as expected and as shown in Fig.5.11(a) where the horizontal dark gray plane indicates the load distribution in a fixed network. For higher mobility users, the load distribution becomes more dependent on the mobility

[3]Without loss of generality, we assume that MVNOs have their own AAA systems. Otherwise, MVNO sessions can be handled like home users for analytical purposes.



(a) Mobility Pattern C (b) Mobility Pattern D

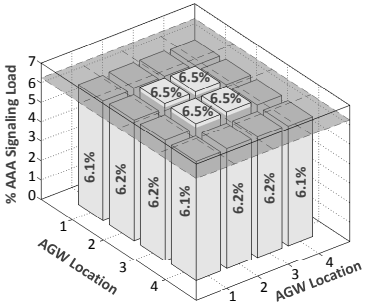
Figure 5.10: Mobility patterns in the 4×4 AGW network (For roaming, each side of the border AGWs lets traffic leave with a likelihood of 25%).

Table 5.4: Parameters for the distributed AAA case study.

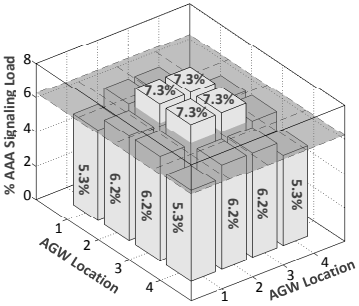
User sessions	Home users: 5 sessions/sec/AGW, Roaming Users: 1% of the home users, MVNO Users: 20% of the home users. Home users in roaming partners networks are equal to roaming users in own network. Percentage on-net and off-net traffic 95% and 5%.
User distribution	Home users & roaming Users: uniformly distributed among all AGWs, MVNO Users: non-uniform (see Fig.5.12)
Session durations	Home & MVNO users: 30 or 60 min, Roaming Users: 10 min
Mobility	Patterns C & D with roaming (see Fig.5.10(b)) , AGW residence time = 58 mins for border AGWs and 29 min for central AGWs. Residence time is gamma distributed with coefficient of variation of 1.2. Only 30% of the users are mobile.
Protocol	Authentication Scheme: CHAP based (1 exchange is assumed, no home agent authentication is used), Interim interval: Half the session duration for each user type. Authorization Lifetime = full session duration for all user types

pattern with central AGWs dominating the rest as in Pattern D as shown in Fig.5.11(b)-5.11(c). The central AGWs tend to generate more signaling as they have lower residence times than the rest. Thus depending on mobility, AAA systems may not be simply planned based on the users' distribution in mobile networks.

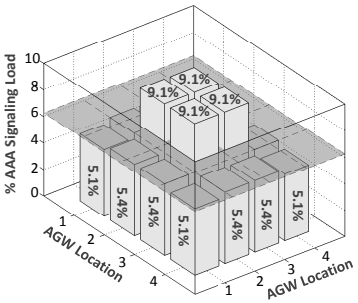
We now proceed to investigate the AAA signaling at each AAA in the network. We assume home, roaming, and MVNO users. The users' distribution, session statistics, mobility profiles, and protocol parameters are given in Table 5.4. For roaming and MVNO users, we consider two cases. The first case assumes that each authentication request is forwarded to the visited network by one of the local AAA systems (i.e., AAAs



(a) % load AAA distribution (Any Pattern, Low Mobility)



(b) % AAA load distribution (Pattern C)



(c) % AAA load distribution (Pattern D)

Figure 5.11: % AAA load distribution in a network of 16 AGWs (Dark plane indicates the signaling in an equivalent fixed network)[E_R=58 and 29 mins for outer and central AGWs resp., C_R=1.2, E_s=60 min, $\Delta_T = E_s/2$, $\Delta_M = E_s$, %Mobility = 30%, %Stationary users = 70%].

1-3) through AAAg. The other case assumes authentication delegation to the local AAA system after the initial authentication. Thus, if the serving AAA system is AAA1 and the mobile hands off from AGW1 to AGW5, AAA1 handles the authentication locally and AAAg is not contacted. On the other hand, if the mobile moves from AGW_{12} to AGW_{11}, AAA_2 contacts the visited network through AAA_g for authorization. Re-authentication and accounting requests are always forwarded to the visited network.

First, let us consider the locally handled signaling load on each AAA system for session durations of 60 and 30 mins as shown in Figs.5.13(a)-5.13(b). For mobility pattern C, the AAA load pertaining to home users, which is locally handled, is a function of the number of AGWs served by the AAA system and hence AAA1 receives the highest signaling load. AAA2 has slightly higher signaling than AAA3 due to the fact that

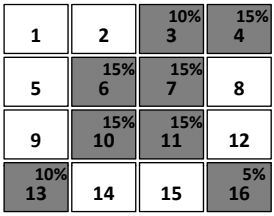


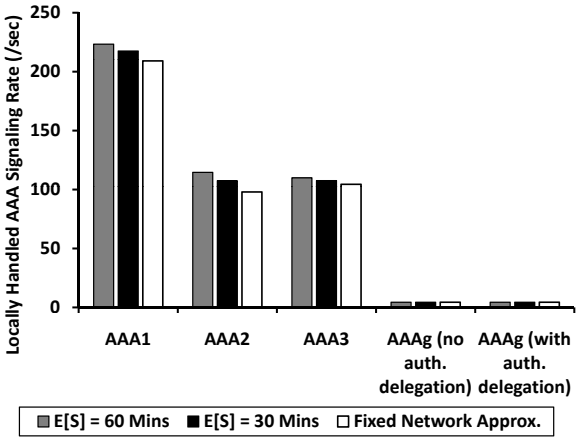
Figure 5.12: User distribution for MVNO users.

AAA2 captures more terminating sessions - AAA2 serves the central four AGWs while AAA3 tends to serve more partial sessions as users may roam outside the network. For mobility pattern D, the signaling load at AAA2 is higher as users are more likely to visit the central AGWs in this pattern. The same trends apply when the session duration is halved. Since the chosen AGW residence times are high with only 30% mobility, the fixed network model offers decent estimates for the signaling load at each AGW. Since AAAg only handles requests from home users roaming in other networks, the load it handles from local requests is insignificant compared to the load handled by the other AAA systems (i.e., AAAs 1-3).

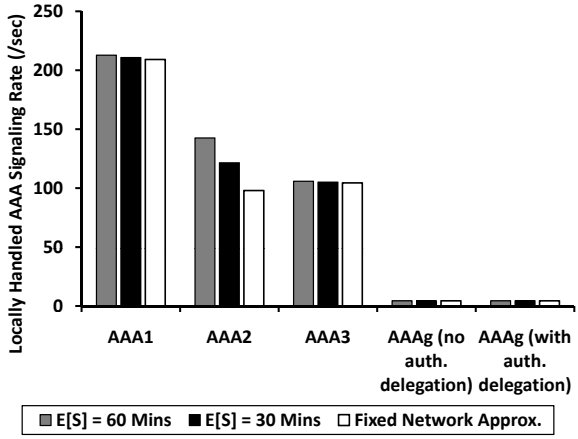
For the proxy AAA signaling in Figs.5.14(a)-5.14(b), AAAg handles the highest load as it acts as a gateway to other networks. The proxy load at AAA2 is higher than the rest for mobility pattern "C" due to the non-uniform user distribution of the MVNO users (see Fig.5.12) while it dominates others for pattern "D" due to the mobility pattern and the non uniform distribution which both favor the central AGWs. We see that the use of authentication delegation from the AAAg to the local AAAs can reduce the signaling load at the gateway to almost 40% of the current load as AAAg needs to only authorize requests when they move between AGWs belonging to different AAA systems. Such results indeed help the designer of AAA systems to select the right server size and properly map AGWs to the AAA systems depending on the generated load.

5.3.4 The Impact of Different Authentication Protocols

Emerging wireless architectures are envisioned to encompass multiple cellular technologies (e.g., LTE and WiFi, EVDO and WiMAX). A design challenge which we consider here is the fact that some AGWs may belong to one cellular technology which may use different authentication mechanisms than the other cellular technologies in the network. For instance, WiMAX AGWs (a.k.a, ASN-GW gateways) use the Extensible Authentication Protocol (EAP) authentication methods which require a significant number of rounds with the client (e.g., 6 rounds). On the other hand, in EVDO systems AGWs (a.k.a., Packet Data Serving Nodes (PDSNs)) use one round Challenge-



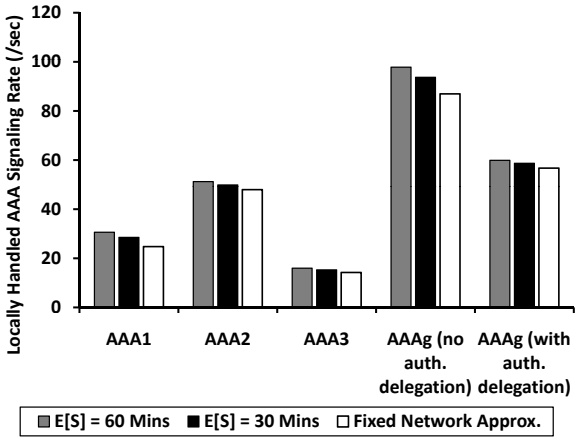
(a) Local AAA signaling load (Pattern C)



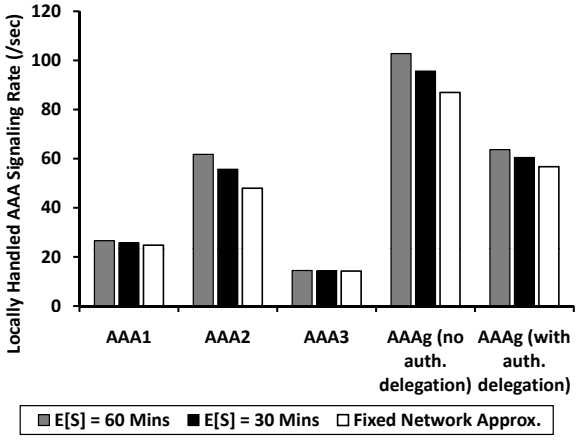
(b) Local AAA signaling load (Pattern D)

Figure 5.13: Percentage AAA load distribution in the network (local signaling).

Handshake Authentication Protocol (CHAP) authentication based schemes. Due to the long duration of EAP authentications, Fast Handoff (FH) mechanisms where proposed to reduce authentication signaling to one round [31] after the initial authentication is performed. Our goal in this discussion is show that even under such challenging design



(a) Proxy AAA signaling load (Pattern C)



(b) Proxy AAA signaling load (Pattern D)

Figure 5.14: Percentage AAA load distribution in the network (proxy signaling).

conditions with different protocol deployments, our generic AAA signaling planning framework is still applicable. Our framework is especially useful as the the residence times for the WiMAX and EVDO AGWs may largely differ due to the users' behavior and wireless coverage differences between both technologies.

Table 5.5: Parameters for EAP and CHAP authentication scenarios.

User sessions	5 sessions/sec/AGW in both WiMAX and EVDO systems.
User distribution	uniformly distributed among all AGWs.
Session durations	60 min.
Mobility	Patterns C & D *without* roaming (see Fig.5.10) , AGW residence time = 58 mins for EVDO AGWs and 29 min for central WiMAX AGWs. Residence time is gamma distributed with coefficient of variation of 1.2.
Protocol	Authentication Scheme: CHAP based (1 exchange is assumed, no home agent authentication is used) for EVDO. For WiMAX EAP authentication based on the TTLS method with 12 AAA authentication messages is assumed [166]. Interim interval: Half the session duration for each user type. Authorization Lifetime = full session duration for all user types.
AAA assignment	See Fig.5.9(b).

Let us consider the case where the central AGWs in Fig.5.10 are WiMAX based while the rest are EVDO AGWs. The parameters we use in this study are summarized in Table 5.5. The results are illustrated in Fig.5.15 and Table 5.6. From Fig.5.15, we see a significant load generated by the WiMAX AGWs compared to the EVDO AAA systems which is inline with the recently discussed challenges in the industry as in [4]. Again, mobility pattern D results in higher signaling load than pattern C since in pattern D users tend to go to the central WiMAX AGW nodes. Furthermore, the results show that the gain from applying EAP fast handoff (FH) schemes [31] in terms of signaling reduction in the vicinity of 10% (in pattern C) to 20% (in pattern D). This indicates that the fast handoff gain is maximized only when users are served by the central AGWs belonging to the WiMAX network and is reduced as the users make vertical handoffs between EVDO and WiMAX networks. Due to the difference in the authentication protocols in the network, the fixed network approximation results in large estimation error for the central WiMAX AGWs. Finally, Table 5.5 shows the signaling load at each AAA system which is simply obtained by summing the signaling loads from the AGWs they serve according to Fig.5.9(b).

Table 5.6: Signaling rate per second at each AAA system.

AAA System	**EAP (Pattern C)**	**EAP (Pattern D)**	**EAP (FH) (Pattern C)**	**EAP (FH) (Pattern D)**
AAA1	236.1	218.8	236.1	218.8
AAA2	363.0	452.9	330.4	364.7
AAA3	114.5	107.7	114.5	107.7

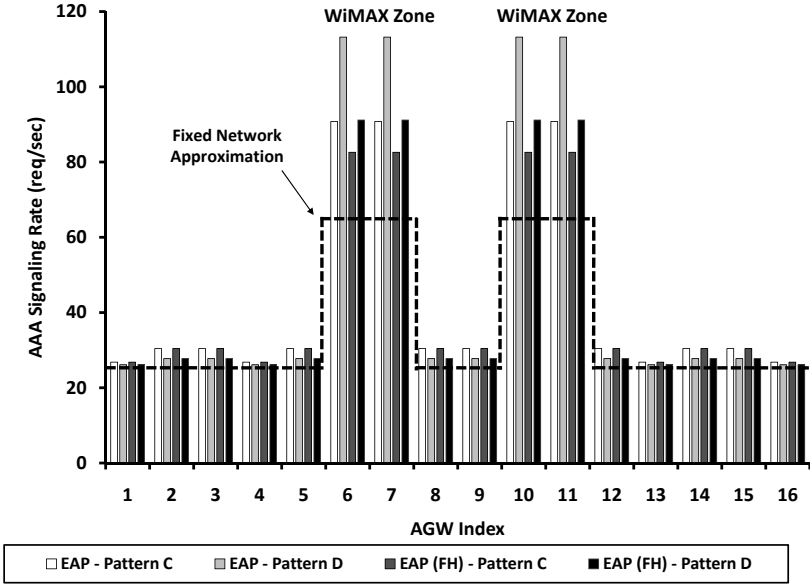
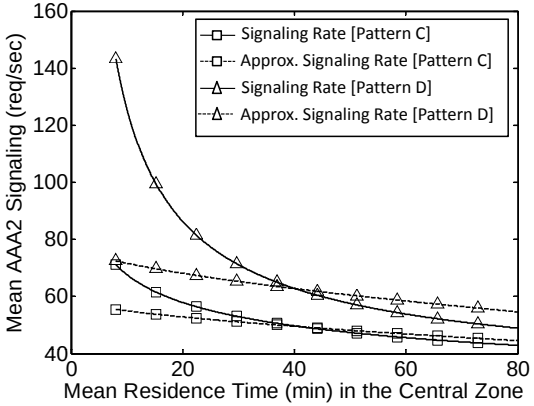


Figure 5.15: The effect of using different authentication protocols [E_R= {58 min for outer AGWs and 29min for central AGWs}, C_R=1.2, E_s=60min, $\Delta_T = E_s/2$,$\Delta_M = E_s$, %Mobility=30%].

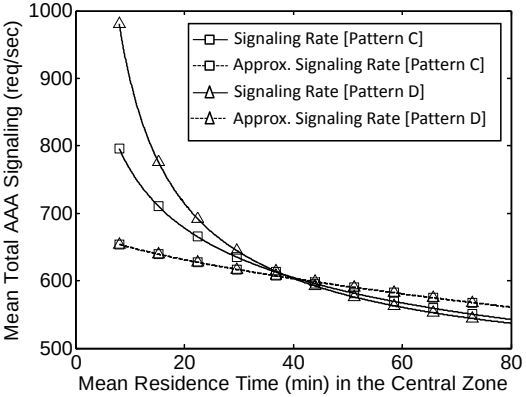
5.3.5 The Impact of the Different AGW Residence Times

The goal of this section is to investigate how the difference in AGW residence times affects the signaling load perceived by the AAA system. As we mentioned earlier, AGW residence times can not be assumed to be identically distributed especially in emerging cellular network architectures which incorporate multiple wireless technologies where cell coverage and user mobility may vary considerably. Our framework relaxes the homogeneous residence time assumption and offers the signaling load using the measured/estimated residence times in the network.

In our analysis, we fix the residence time at the border AGWs in Fig.5.9 to an arbitrary value of 40 mins and vary the residence time at the central AGWs (i.e., AGWs 6,7,10,11). We assume 100% mobility in the network in order to isolate the effect of the difference in the AGW residence times. In our analysis, we investigate the signaling rate in the distributed AAA system in Fig.5.9 and also see the effects on an equivalent centralized AAA system serving all 16 AGWs in the network. We compare the signaling rate calculated with different residence times and that obtained using the average



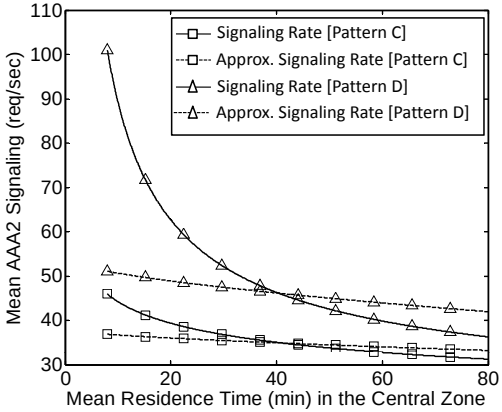
(a) Impact of the difference in residence times at AAA2



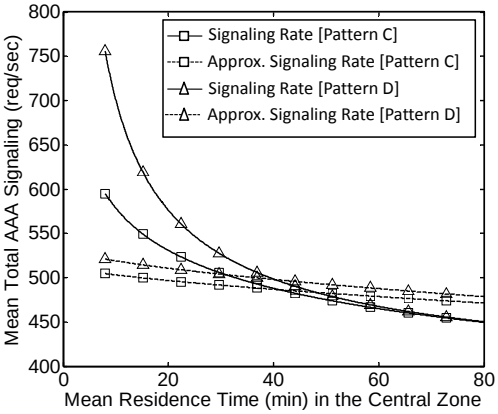
(b) Impact of the difference in residence times on an equivalent centralized AAA System

Figure 5.16: The impact of the difference in residence times in distributed and centralized AAA system deployments (no roaming) [E_R= 40 min for outer AGWs, C_R=1.2, E_s=40min, $\Delta_T = E_s/2$,$\Delta_M = E_s$, %Mobility=100%].

AGW residence time for all AGWs in the network. The latter represents homogeneous residence time in the network and is a commonly used simplifying assumption in many studies in the literature such as in [17, 103, 144]. Our target is to show how the AAA



(a) Impact of the difference in residence times at AAA2



(b) Impact of the difference in residence times on an equivalent centralized AAA System

Figure 5.17: The impact of the difference in residence times in distributed and centralized AAA system deployments (with roaming) [E_R= 40 min for outer AGWs, C_R=1.2, E_s=40min, $\Delta_T = E_s/2$,$\Delta_M = E_s$, %Mobility=100%].

signaling rate estimates depend on the residence time as function of mobility pattern and roaming. Up to our knowledge, such results can only be obtained by simulations.

Our results are summarized in Fig. 5.17. We show the signaling load at AAA2 in the distributed AAA system in Fig.5.9 and compare it with the equivalent centralized system. The signaling loads at (AAA1 and AAA3) follow similar trends and are not shown for brevity. In Figs.5.16(a)-5.16(b), we assume that users are not allowed to roam outside the network under consideration while roaming is possible in Figs.5.17(a)-5.17(b). As shown in Fig.5.16(a), we see that the signaling rate obtained from both mobility patterns C and D and that obtained by the homogeneous AGW residence time approximation (the dashed line) intersect when the residence time in the central zone is equal to the residence time in the border AGWs (i.e., at mean AGW residence time of 40 mins). In addition, we see that the difference between the actual signaling load and its approximation is less severe when the mobility is low (i.e., for residence times of 40 mins and above). Again, mobility pattern D results in higher signaling than pattern C due to the fact that pattern D favors the central region. In Fig.5.16(b), we investigate the impact of the different residence times on the signaling rate received by an equivalent centralized AAA system. The most important observation is that the mobility pattern now plays a role in determining the total signaling load even when no roaming exists. Assuming a homogeneous residence time can result in large errors when mobility is high (i.e., mean residence times less than 40 mins). In Figs.5.17(a)-5.17(b), we investigate the signaling rate when roaming is possible to other networks. Although the trend is quite similar to that in Figs.5.16(a)-5.16(b), the difference between the signaling rates due to the two mobility patterns is larger as a consequence of roaming.

To sum up, we conclude that the homogeneous residence time assumption linearizes a non-linear effect on the signaling load in both roaming and non roaming scenarios. Such linear approximation can be acceptable for low mobility scenarios and depending on the mobility pattern can lead to large errors when mobility is high. In addition, we also showed that the AAA signaling rate depends on the mobility pattern not only when there is roaming but also when the residence times in the network are different.

5.3.6 Summary of Planning Methods and Their Applicability

Although the general distributed AAA model in Section 3.6 is sufficiently generic to cover various practically relevant scenarios, the basic AAA model in 3.5 as well as the fixed networks AAA model in Section 3.4 offer significantly easier ways of obtaining AAA signaling load in relatively simple AAA deployments. Table 5.7 shows the applicability scope for the three models. Note that in low mobility scenarios the fixed and basic models are in fact able to capture the signaling rate in distributed AAA systems or in cases with different authentication protocols as users do not move between AGW regions. Hence, signaling from each protocol and from each AGW can be scaled according to the session arrival rates at each AGW. Due to mobility pattern and roaming effects the generalized model should be generally used for distributed AAA systems in high mobility scenarios.

Table 5.7: Summary of the planning models and their applicability to a range of scenarios.

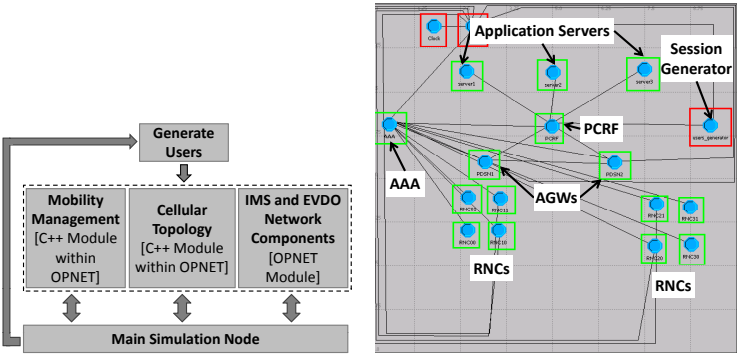
		Model	Different Auth. Protocols	Interim Interval & Auth. Lifetime	Auth. Delegation	EAP Fast Handoffs	General Session Distribution	Roaming	Different AGW Residence Times	Mobility Pattern	Number of Services	AGW Signaling Load Distribution	Stats for Session Holding Time Types
Centralized AAA Deployment	Low Mobility	Fixed Model	A	✔	x	x	✔	-	-	x	✔	A	-
		Basic Mobile Model	A	✔	✔	✔	-	-	A	x	✔	A	A
		Generalized Model	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔
	High Mobility	Fixed Model	-	-	-	-	-	-	-	-	-	-	-
		Basic Mobile Model	-	✔	✔	✔	-	-	-	-	✔	-	A
		Generalized Model	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔
Distributed AAA Deployment	Low Mobility	Fixed Model	A	A	x	x	A	-	-	x	A	A	-
		Basic Mobile Model	A	A	A	A	-	-	A	x	A	A	A
		Generalized Model	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔
	High Mobility	Fixed Model	-	-	-	-	-	-	-	-	-	-	-
		Basic Mobile Model	-	✔	-	-	-	-	A	-	A	-	-
		Generalized Model	✔	✔	✔	✔	-	✔	✔	✔	✔	✔	✔

Acronyms

✔ : This aspect is supported by the model	A: This aspect is approximated by the model
X : This aspect is minorly or not impacting	- : This aspect is not supported by the model

5.4 Handoff Delay Optimization in Multi-Service Mobile Networks

The focus of this section is on the performance of the proposed authentication delay minimization mechanism described in Section 4.3 and its signaling load in the network.



(a) Simulation model with OPNET (b) OPNET screen shot for the simulator

Figure 5.18: Simulation model for the delay optimization logic

In our evaluation, we developed C++ modules within the OPNET simulator. We assumed an EVDO network with IMS capabilities. As shown in Fig.5.18(a), we see that our simulator incorporates three modules: one that represents IMS and EVDO network components and handles their respective signaling, another module that represents the cellular topology in terms of cells and their mapping to radio network controllers, and the last handling the users' mobility including the radio signal attenuation models. The first module is programmed within the OPNET Modeler using its finite state machine description while the other two are written in C++ modules which are called from within OPNET. In depth details of our OPNET simulation environment is available in [140].

To study the effect of our mechanism on both the signaling plane and the data plane, we generate a Poissonian load of sessions with Lognormally distributed durations and we randomly select one of the generated sessions to act as probe user. The probe user generates VoIP traffic and we measure the number of dropped packets as we discussed in Section 4.3. We simulate the signaling mechanism for two neighboring AGWs each supporting four RNCs. Each RNC serves $N \times N$ cells. We assume a composed service running over three application servers and assume that authorization signaling from the PCRF is forked to all application servers at the same time. The OPNET simulation model is shown in Fig.5.18(b). The application servers service time is modeled using M/M/1 behavior. Table 5.8 lists the fixed parameters of the VoIP application, the wireless channel, and the topology used in our simulation. All simulation results are conducted within the 90% confidence levels.

To study different mobility patterns, we modify the well-known Waypoint model by limiting the distance a node travels during each movement epoch. We refer to this dis-

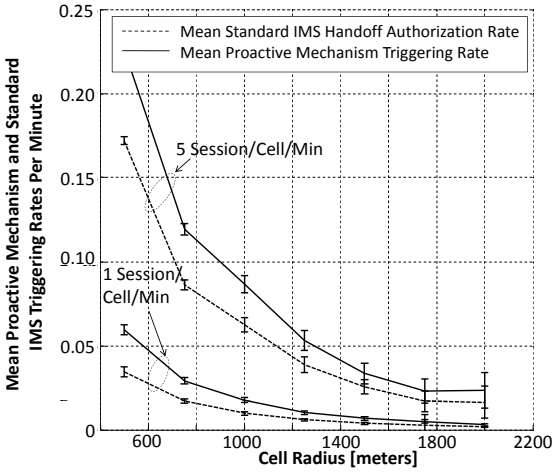
tance as the *span*. In other words, after pausing the node picks a random destination that is utmost *span* distance units away. By controlling the *span* variable in our simulator we are able to simulate different mobility patterns during each session duration. For instance, long spans result in long straight movements or highly directional mobility, whereas short spans result in localized movement or highly random mobility. We also consider the corrections suggested by [167] by having the minimum speed > 0. Note that we do not incur the known issue that users in Waypoint simulations tend to go towards the center *after a long time* as in our simulations session durations are limited and new sessions are randomly placed in the coverage area and are removed once they finish. It is also noteworthy to state that our choice of the Waypoint model rather than the statistical approaches using the residence time to verify the operation of our mechanism is because we would like to test the performance of the prediction scheme. The waypoint allows tracing each user and performing relevant mechanisms on a per user basis which is non-trivial using statistical methods of residence time.

Table 5.8: Simulation parameters.

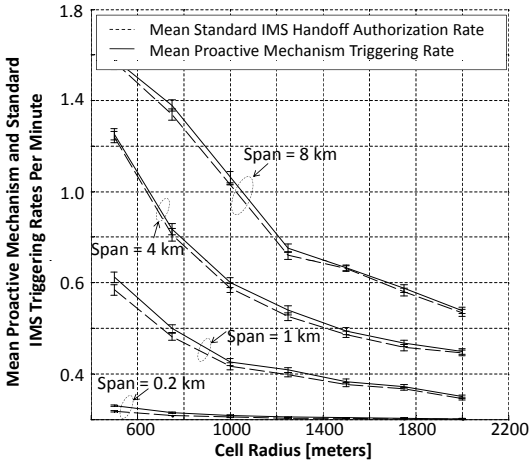
Aspect	Parameters
VoIP	Session duration:10 min, Frame duration, Talk Spurt, Silence Period (ms):(20,6.5,6.5), Codec:EVRC, coding rate:8.5kbps.
Wireless Channel	Mobile Node's Tx power 250 mW, Freq.=1.9 Ghz, channel rate = 38.4 kbps, channel power loss exponent = 3.7, LogNormal shadowing standard deviation = 4.1.
Topology	2 AGW areas, 2×2 RNCs per AGW, 3 Application. Servers
Link Delays	RNC-AAA: 40ms, RNC-AGW: 40ms, PCRF-AAA: 100ms, PCRF-AGW: 100ms, PCRF-AS: 200ms.

Figure 5.19 summarizes the simulation results relevant to the performance of our mechanism. Results relevant to the operation of the mechanism and its ability to mitigate delay is demonstrated for a VoIP application are shown in Section 4.3. We first study the effect of the cell radius on the mean triggering rates of our proactive signaling as well as the standard IMS authorization schemes, as described in Section 2.4.3. We simulate two session arrival rates (i.e., 1 and 5 sessions/cell/min). As shown in Fig.5.19(a), as the cell radius increases, the mean signaling rates for both schemes decrease due to the reduced likelihood of AGW handoffs. We also observe that the proactive signaling rate is larger than that of the standard IMS procedure. This is due to the fact that the standard IMS signaling is triggered once per handoff while the proactive mechanism incurs extra triggers due to the mis-predicted handoffs (i.e., false alarms). We conclude that even though the AGW handoff rate increases significantly as the cell size is decreased, the proposed mechanism signaling scales similarly to the IMS standard procedure.

In Fig.5.19(b), we study the effect of the mobility pattern on the mean signaling rates to obtain deeper insights on the signaling overhead due to false alarms. We perform simulations for various Waypoint mobility spans ranging from highly random (0.2 km) and highly directional (8km) movement patterns. We observe that the triggering rates

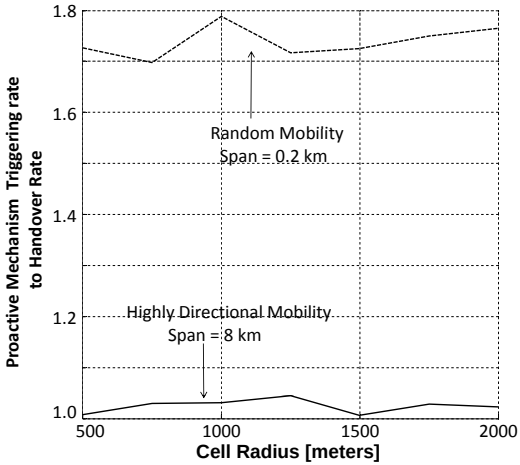


(a) Mean mechanism triggering & handoff rates vs. cell size (Span=0.2km)

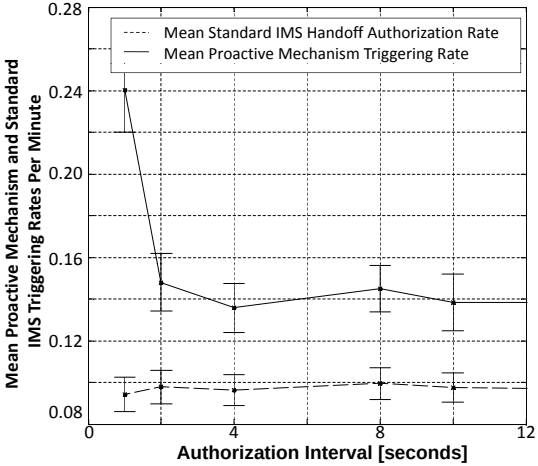


(b) Mean mechanism triggering & handoff rates vs. cell size for different mobility patterns

Figure 5.19: Simulation parameters common to all figures (adapted from [136]) [5x5 cells/RNC, Authorization Interval = 150 sec, AS Load=50%, δ set at 90% AS loading, 1 session/cell/min].



(c) Number of mechanism executions per handoff



(d) Mechanism triggering rate as a function of the authorization interval duration

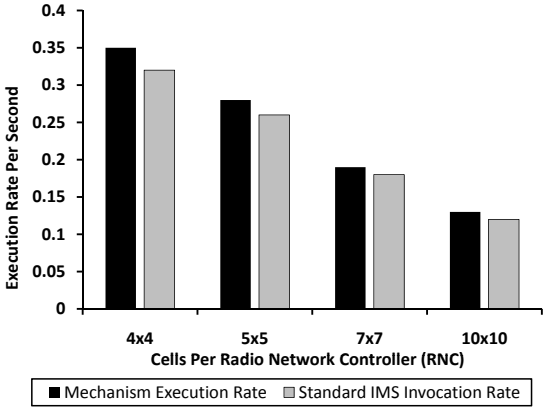
Figure 5.19: Simulation parameters common to all figures (adapted from [136]) [5x5 cells/RNC, Authorization Interval = 150 sec, AS Load=50%, δ set at 90% AS loading, 1 session/cell/min].

for both standard IMS and the proactive mechanism are very similar and differ the most for random movers, albeit at low rates. For random movers, the triggering rates are very low as they barely leave their initial access gateway due to the frequent changes in direction during their sessions. On the other hand, highly directional movers (i.e., high span) are more likely to leave their gateway resulting in larger signaling rates. For instance, the signaling rate for a cell radius of 0.5km and a span of 8km is approximately 16 times that for spans of 0.2km. Although the mobility pattern highly affects the resulting signaling rate, the proposed mechanism scales similarly to the IMS signaling scheme. It is noteworthy to state that any optimization mechanisms for signaling reduction (e.g., delegation and hierarchical design as in [9, 32]) can be used without fundamental changes.

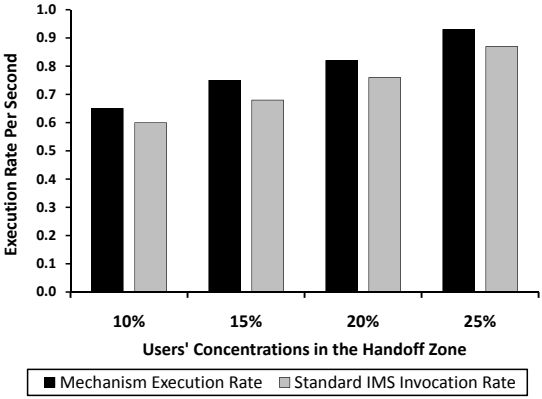
In order to obtain bounds on the performance of the proposed mechanism, we study the ratio of the proactive signaling mechanism to the handoff rate for random and highly directional movement patterns as shown in Fig.5.20(c). We observe that for highly directional movers the proactive triggering mechanism is executed almost at the same rate as the handoff rate (i.e., the standard IMS mechanism). On the other hand, proactive signaling is triggered at approximately double the handoff rate for random movers. This is due to the fact that random movers result in a large number of false alarms and hence larger number of proactive signaling executions if they are near the border. As such, Fig.5.20(c) establishes performance bounds using the two extreme mobility patterns and hence we expect that in the worst case the triggering rate of our proactive mechanism is twice that of the standard IMS scheme.

An important setting for our mechanism is the authorization interval (see Mechanism 3). This interval needs to be set low enough in order not to waste resources on the target AGW and RNC and high enough to avoid excessive invocation of the mechanism while users move in the handoff zone. In Fig.5.20(d), we observe that mean triggering rate of our mechanism is barely affected for authorization intervals as low as 2 seconds below which the signaling rate increases rapidly. This means that mean residence time for a moving user in the triggering zone is approximately 2 sec and hence the mean authorization interval must be selected to be greater than 2 sec to avoid excessive transmission of HI messages. However, one should not set this interval around 2 sec in practice as real users' behavior may vary and hence higher values should be used (e.g., 30 sec or more).

We now attempt to verify the mechanism's scalability as function of the number of cells per RNC as well as the users' concentration in the handoff zone. Let us first investigate the impact of the number of cells/per RNC on the performance of our mechanism in comparison with the standard IMS mechanism. Since the handoff rate between RNC regions is inversely proportional to the square root of the number of cells they serve [144], it is expected that the IMS standard mechanism to follow this trend as it is only triggered during handoffs. Our goal is to verify that our mechanism also scales similarly and following a square root relationship. From Fig.5.20(a), we first observe that as the number of cells is increased, the signaling rates for our mechanism and the standard IMS mechanism decrease as it takes longer to leave a larger region leading to less handoffs.



(a) Impact of the number of cells/RNC



(b) Impact of the number of the users' concentrations in the handoff zone (5x5 cells/RNC)

Figure 5.20: The mechanism scalability as a function of the number of cells per RNC and users' concentration in the handoff zones. Common simulation parameters [1 session/cell/min, Session duration = 10 min, Span = 1km, cell radius = 1km].

By performing linear least square fit for the signaling rates versus the number of cells per RNC (i.e., 16, 25, etc), we get $\frac{1.48}{\sqrt{x}} - 0.02$ or complexity of $O(\frac{1}{\sqrt{x}})$ as expected. On the other hand, the fit for the invocation rate of the proposed scheme is given as

$\frac{1.34}{\sqrt{x}} - 0.01$ with root mean square error (RMSE) of 0.4% or complexity of $O(\frac{1}{\sqrt{x}})$. We conclude that our mechanism scales similarly to the IMS mechanism as the number of cells per RNC is changed. From Figure 5.20(b), we also see that our mechanism scales similarly as the standard IMS mechanism even when the users' concentration in the handoff zone is relatively high (i.e., 25%).

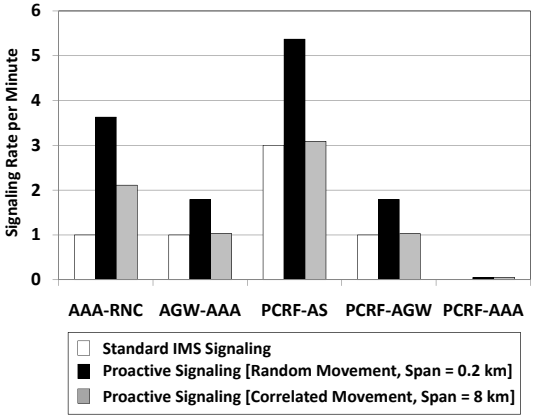


Figure 5.21: Signaling rate for the standard IMS and the proactive signaling mechanisms.

So far we have only investigated the mechanism invocation rate in general, we now investigate the signaling rate in relevant interfaces (i.e., AAA-RNC, PCRF-AAA, PCRF-AS, and PCRF-AGW) for the two sample mobility patterns. From Fig.5.21, we observe that for highly directional movers, the signaling load of the proactive and the standard IMS mechanisms are almost the same on all interfaces except on the AAA-RNC interface. This is because of the newly introduced HI messages in our proactive scheme. Notice that since the SALI and the HNR messages are only triggered when there is a major change in the application server loading, they barely result in any additional loading on the PCRF-AAA and the AAA-RNC interfaces. For random movers and due to the large likelihood of false alarms the signaling rate for the proactive mechanism relative to the standard IMS mechanism is approximately three-folds on the AAA-RNC interface due to the more frequent transmission of the HI messages and almost two-folds on rest of the interfaces. In real deployments, a spectrum of movement patterns is likely the case and hence the observed signaling rates on the corresponding network interfaces will be in between that of random and directional movers.

To sum up, from the results in Figs.5.19-5.21, we conclude that our mechanism does not impose a significant load on the serving cellular network and scales similarly to the standard IMS mechanism as function of various parameters including cell size, cells per RNC, users' concentration in handoff zones, session arrival rates, and mobility patterns.

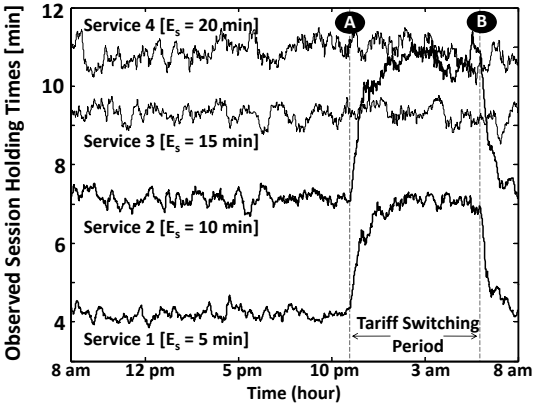
5.5 Optimizing Accounting in Multi-Service Mobile Networks

In this section, we examine the operation of our accounting optimization mechanism in Section 4.4 under a wide range of operational conditions. Our objective is to show that the mechanism is robust to various mobility and session conditions, after demonstrating the basic operation of our proposed mechanism for fixed network deployments in Section 4.4.6. In this section, we study realistic scenarios of mobile networks under conditions of variable loads, tariff switching, failovers, and roaming scenarios. We conclude our discussion by examining the execution delay and the rate of invocation of the proposed mechanism to demonstrate that it is lightweight and easy to implement.

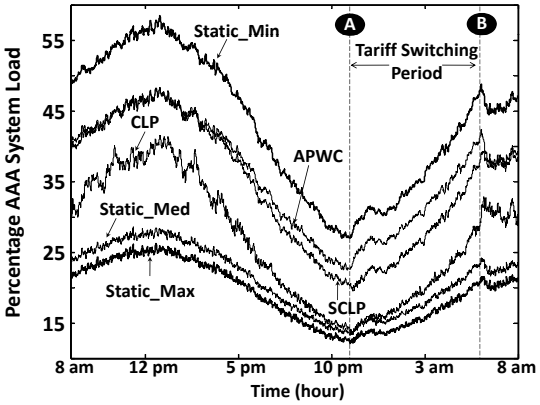
In our simulations, we assume without loss of generality that the analyzed network is an area composed of 3×3 NASes. NAS coverage areas are different and the movement between their areas is assumed to be random. As users move between NAS regions, they can randomly trigger any of the four possible session scenarios (see Table 4.2). As in Section 4.4.6, we validate our mechanism's performance by comparing its operation with three policies with static interim interval settings, to mimic current systems, i.e., Static_Min, Static_Med, and Static_Max. The interim settings for Static_Min are set to 1 min for all services. For Static_Med and Static_Max policies, the interim settings are fixed to half and full mean session durations respectively. The interim intervals are updated by invoking the optimizer when session and/or arrival statistics change by 5% and only after a grace period of 75 seconds has passed since the last update.

5.5.1 Impact of Mobility

Our goal is to investigate the benefits of our policies and stability in maintaining the expected behavior in mobile environments and under complex scenarios with multiple services characterized by different session durations and tariffs. In order to demonstrate that our approach works properly in non centralized AAA deployments, in our simulations, the AAA system under consideration only receives accounting reports from the central NAS while other NASes report to other AAA systems. The central NAS has a mean residence time of 25 min. All NASes serve four services with mean durations of 5,10,15, and 20 mins and service rates of 0.1, 1, 0, 0.02 price units/min. The zero cost is used to indicate that service 3 is a flat rate service and to investigate the effect of service tariffs on the behavior of our schemes. Further investigation of more complex pricing plans can be carried as in [54] and is a future research item for this thesis. For comparison purposes we assume that the arrival rates of all services are the same. Tariff switching is applied to services 1 and 2 between 11pm and 6am (see Fig.5.22(a), instants A and B). During this time, the service costs are halved and the session durations double from 5 and 10 to 10 and 20 mins respectively. For all services, the mean load varies during the day following a sinusoidal pattern.



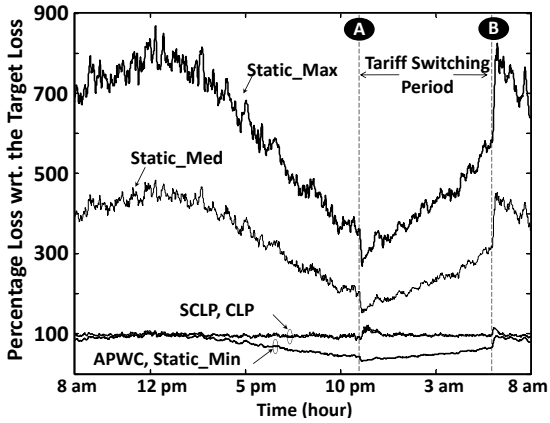
(a) Mean Session Holding Time Observed by NAS



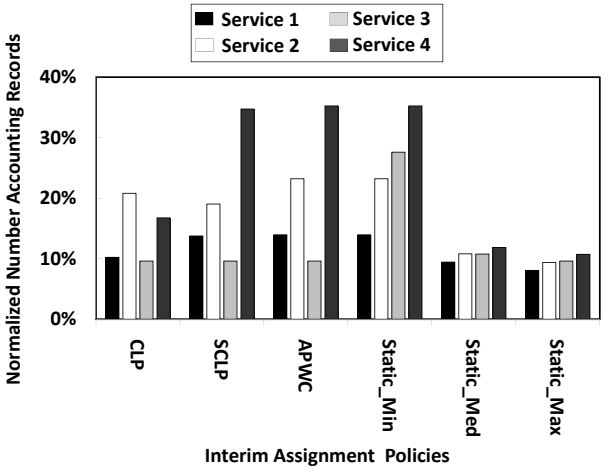
(b) AAA System Load

Figure 5.22: System's performance in a mobile network environment (adapted from [145]) [tariff switching occurs between from 7pm-7am, λ_i=1+0.4sin($\frac{2\pi}{24hr}t$)/s, S/CLP target loss(L_1^{max}) = 500 units, AAA capacity P = 150 req/s, average window sizes = 100, mean AGW residence times are {10, 22, 23; 43, 25, 10; 17, 10, 11.6} min, 30 indep. simulation runs, 4 hr warm up period, 95% confidence (change within 3% variation)].

- *The session holding times (Figs.5.22(a)):* In our method, the mean session holding time is estimated as the weighted average of the mean holding time from all mobility components as E_{s_i}=$\frac{\sum_x \lambda_i^{(x)} E_{s_i}^{(x)}}{\sum_x \lambda_i^{(x)}}$,$x \in \{F, O, Tr, T\}$. To validate the cor-



(c) Potential Loss



(d) Number of accounting records per day

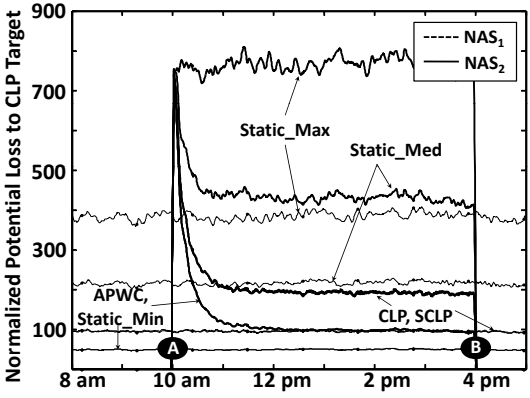
Figure 5.22: System's performance in a mobile network environment (adapted from [145]) [tariff switching occurs between from 7pm-7am, λ_i=1+0.4sin($\frac{2\pi}{24hr}t$)/s, S/CLP target loss(L_1^{max}) = 500 units, AAA capacity P = 150 req/s, average window sizes = 100, mean AGW residence times are {10, 22, 23; 43, 25, 10; 17, 10, 11.6} min, 30 indep. simulation runs, 4 hr warm up period, 95% confidence (change within 3% variation)].

rectness of this method, we compare our estimated session holding time from the four components to the theoretical mean channel holding time from [37] which assumes pre-knowledge of the session duration S_i and the NAS residence time, R, (i.e., the time a mobile device spends in the NAS coverage area). In short, [37] models the mean channel holding time by the minimum of the whole session duration S_i and the residence time in the NAS region, R and hence the mean session holding time is given as $\frac{E[R]E[S_i]}{E[S_i]+E[R]}$ under exponential distributions assumptions. Notice that we can not directly use such model because estimating R is hard in realtime and the knowledge of S_i requires communications between all AAA systems which serve the mobile sessions. As shown in Fig.5.22(a), due to mobility, we observe that the estimated session holding times for all services are less than their mean values. Our holding time estimate matches the theoretical estimates in [37] (e.g., $\frac{5*25}{5+25} = 4.2$ and $\frac{20*25}{20+25} = 11.1$ for services 1 and 4 resp.). This confirms the feasibility of our mobility estimation method using four components.

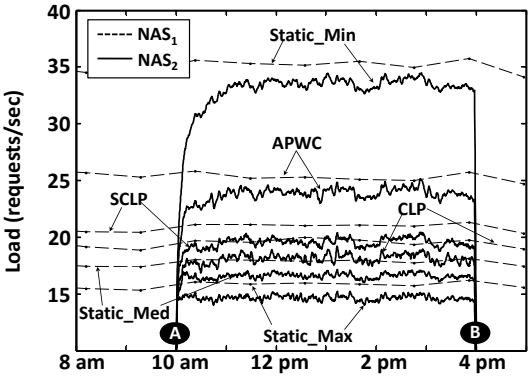
- *The AAA system load and potential loss behavior (Fig.5.22(b)-5.23(c)):* In this case, we see similar trends for the static and optimization policies as in the fixed network case we showed in Chapter 4 in Figs.4.10(b)-Fig.4.10(c). This verifies the proper and consistent operation of our schemes in mobile environments. In this regard, the CLP offers more optimal load performance than the SCLP while both maintain the same loss target. The loss from the APWC mechanism is identical to the Static_Min policy while they differ in the AAA system load. This is because Service 3 has zero cost which is considered by the APWC while ignored by the Static_Min policy.

- *The mean number of interims (Fig.5.23(d)):* We also observe similar trends as in the fixed network example in Chapter 4 in Fig.4.10(d). However, the effect of the service tariff is reflected on the mean number of interims generated by the proposed policies. Common to all of our optimization mechanisms the number of interims for service 2 is relatively large and that for service 3 is low which reflects their relative tariffs. We also see that the APWC produces the same number of interims as the Static_Min policy for all services except for service 3 due to its cost and thus explains the load difference between the APWC and the Static_Min policy in Fig.5.22(b).

From our results in Fig.5.22 for mobile networks as well as the basic results in Fig.4.10 for fixed networks, we confirmed that our policies allow much better control of the potential loss relative to the static policies and are more resilient to changes in session statistics as they manage to either minimize the loss (i.e., in the APWC case) or maintain a constant loss target (i.e., in the CLP and SCLP cases) in fixed and mobile networks.



(a) Normalized Potential Loss



(b) AAA System Load from NASes 1 and 2

Figure 5.23: Failover effect (from [145]), λ_i=1/s, S/CLP target loss = 700 and 1400 for NAS_1 and NAS_2 units, AGW residence times are $\{40, 60\}$ min for NAS_1 and NAS_2 resp., AAA capacity P = 80 req/s, 30 indep. simulation runs, 4 hr warm up period, 95% confidence (change within 3% variation) [dashed lines are used to represent slightly fluctuating curves in (a)-(b) for clarity].

5.5.2 Impact of NAS Failovers

Let us now investigate the mechanism's behavior when another NAS fails over to the AAA system under consideration. In this regard, we study a mobile network configuration where the AAA system normally serves one NAS, which we refer to as NAS1, and a new NAS (i.e., NAS2) fails over to the AAA system under consideration after

its serving AAA fails. The NAS sizes are assumed to be different with NAS2 covering a larger area. In order to clearly see the transient behavior of the policies, we study the system under constant load and we assume that the original AAA server for NAS2 stopped responding due to overload. We assume that NAS2 was always instructed to have the maximum allowable interim interval setting at $\mathbf{\Delta_T^{max}}$ prior to fail over and hence resulting in the largest possible potential loss at the fail over event. Each NAS serves three services with equal arrival rates but with different tariffs. The service tariffs for services 1 to 3 from NAS1 are 0.2, 1, 0 price units and for services 4 to 6 from NAS2 are 0.4, 2, 0 price units. For comparison purposes, we set the loss targets for the CLP and the SCLP policies such that the potential loss of NAS2 is double that of NAS1. The simulation results for the potential loss and the mean AAA system load are shown in Fig.5.23(a) and Fig.5.23(b).

When NAS2 fails over (i.e., at instant A), then depending on the interim interval settings returned for new sessions coming from NAS2 by the AAA under consideration, a transient behavior of the loss may occur (see Fig.5.23(a)). Since there is no change in the interim setting for the Static_Max policy, no transient behavior is observed for the loss or for the load. Due to the change of the interim interval, all other policies incur a transient behavior. The transient effects in the load curves in Fig.5.23(b) are not as significant as in the case of the potential loss. This in fact shows that changing the interim interval for an operational system does not impact the load drastically while it can majorally change the loss behavior depending on the service costs. Note that for a range of only 10% extra load our optimization policies are able to reduce the potential loss incurred by the Static_Max policy by about 400%.

We also observe that the CLP and the SCLP methods maintain the loss targets for NAS1 and NAS2 at the 100% and 200% levels as shown in Fig.5.23(a) (200%/100% is (2+0.4)/(1+0.2)). The load for the CLP and the SCLP policies from both NASes in Fig.5.23(b) is very similar due to the fact that the tariff targets are proportional to the total service costs. For the APWC policy, the same loss behavior is observed as in the Static_Min policy while the load behavior is observed to be different as the APWC sets the interim settings for services 3 and 6 at $\mathbf{\Delta_T^{max}}$. For all policies, both NASes are jointly optimized and interims are generated to either minimize the loss from both NASes (i.e., the APWC) or to control the loss at the given targets as in the CLP scheme. The slight difference between the loads of NAS1 and NAS2 in Fig.5.23(b) is due to the difference between the NAS sizes where NAS2 poses lower load on the AAA system.

5.5.3 Impact of Roaming Users (Proxy chains)

In some cases such as in roaming, the AAA system connected to the NAS may forward requests to the destination AAA system through few intermediate AAA proxies [46, 91]. This configuration is referred to as the AAA proxy chain. As a result, the optimization carried by one AAA system might be in conflict with the other AAA systems

in the AAA proxy chain. For instance, consider the case for roaming users where NAS_v reports the usage to AAA_v of the visited network which proxies the accounting reports to the home network's AAA_h system. System overload may occur if the optimization is carried out by either of the AAA systems without considering the other. To address this case, when the first request for roaming users is received by the system, a pre-configured capacity Q is requested for the reserved stream from all servers in the chain by AAA_v using an access request message. If the requested capacity is approved by all systems in the proxy chain, then the request is accepted otherwise a reject message is generated. Only one AAA system in the chain (e.g., AAA_v) optimizes the reporting intervals within the prescribed reserved capacity Q while the other AAA systems (i.e., AAA_b and AAA_h) treat these services as non-optimizable. In this case, our policies are left intact with the simple modification to include constraints that limit the load due to the proxied signaling messages below the preconfigured/negotiated limit Q (i.e., $\zeta(\mathbf{\Delta_T}_{px}) < Q$ where $\mathbf{\Delta_T}_{px}$ denotes the interim intervals of the proxied services from network x). The available capacity for local requests is reduced as $(P-Q)$. This simple pre-reservation scheme is suitable for proxy chain configurations as roaming traffic is expected to be low compared to home users' traffic.

Let us now consider a typical configuration which supports roaming users. In this regard, the NAS in the roaming partner's network (which we call here as $NAS_{visited}$) is connected to an AAA system in the visited network. The visited AAA system forwards the accounting traffic to the home AAA system which also supports requests from home NASes (NAS 1 and NAS 2). The visited NAS supports two services each with 1 unit cost to reflect roaming charges while NAS1 serves three services with 0.2, 1, 0 price units/min and NAS2 serves another set of services with price units of 0.2, 2, 0. Before the exchange, both systems negotiate the allocated capacity ($Q = 20$ req/sec to roaming traffic) for the forwarded (proxy) traffic and thus both AAA systems dedicate a maximum load. The visited AAA optimizes the interim values while the home AAA system treats the traffic as non-optimizable. Same results are observed when this is reversed. As shown in Table 5.9, the mean loss is around the target loss limit for the three NASes when using the CLP and the SCLP policies. The load of the $NAS_{visited}$ is below the limit. We also observe that all policies offer significantly lower loss for all NASes without significant load requirements compared to the static policy $\mathbf{\Delta}_{\mathbf{T}}^{\max}$.

Table 5.9: Percentage load and losses for two NASes (i.e., NAS1 and NAS2) and a proxy (adapted from [145]) [30 runs, 95% confidence with error in loss and load below 3% variation, load from NAS1 and NAS2 services is 1/s while that from $NAS_{visited}$ is 0.1/s, P=300 req/s].

	CLP		**SCLP**		**APWC**		**Static_Max**	
	Loss	*Load*	*Loss*	*Load*	*Loss*	*Load*	*Loss*	*Load*
$\mathbf{NAS}_1$	93.5	16.9	94.8	18.5	59.0	19.3	325	14.5
$\mathbf{NAS}_2$	95.9	19.8	97.0	21.4	86.7	20.6	650	14.3
$\mathbf{NAS}_{visited}$	96.6	1.8	98.7	1.8	124	1.7	434	1.2

5.5.4 Computational Performance

Now that we have verified that our mechanism optimizes accounting reliability in mobile networks during failovers and for roaming users, we further investigate the mechanism's performance to demonstrate that it does not require too frequent updates of the accounting interim intervals. In this regard, we study performance in terms of the required execution time for the optimization operation and the number of mechanism's invocations as a function of the trigger setting. The trigger setting is defined as the amount of change in the load and session statistics for services which trigger updating the current interim settings. This in fact determines the mean duty cycle of the mechanism invocation (i.e., the interim intervals update rate) and should always be larger than the mechanism execution delay. In our study cases, we used a standard desktop machine (Intel Core 2 CPU E6700, 2GB of memory, Windows XP OS). In order to observe the effect of the APWC knob parameter, ρ_0, in (4.11) (set at 60%) as we did in Fig.4.10(b) in Section 4.4.6, we study the execution time using two AAA capacities (Case A: 210 req/sec and Case B: 300 req/sec) to reflect two different system utilizations. As shown in Table 5.10, we observe very low execution times for the SCLP method compared to the CLP and the APWC methods. We also observe that system utilization (compare Case A and Case B) barely affects the performance of the CLP and the slight difference is rather within the confidence limits of the test. On the other hand, the performance of the APWC scheme is affected with the system load (e.g., compare Cases A and B for 12 and 15 services). This is due to the fact that the system load exceeds the APWC knob setting and hence the non-linear weight function W in (4.11) starts to have significant values in the objective function and thus impacts the optimization time. We conclude that due to the superior performance of the SCLP scheme, it is possible to directly implement it into the AAA servers as a simple module as in [129, 130, 168].

Table 5.10: Mechanism Execution Delay (ms) (adapted from [145]) [All results are within 50 ms for APWC and CLP and within 5 ms for the SCLP scheme with 95% confidence using the mean batch method, 30 batches, constant unit load from all services].

	No. Services	**2**	**4**	**8**	**12**	**15**
Case A	**APWC**	763	791	857	1618	1919
	CLP	818	895	864	1009	1265
	SCLP	3	5	9	12	16
Case B	**APWC**	760	773	858	994	1698
	CLP	824	899	869	997	1194
	SCLP	2	4	8	11	14

Let us now investigate the effect of the mechanism triggering threshold on the execution rate. As shown in Fig.5.24, we observe that increasing the optimization triggering threshold drastically reduces the mechanism's invocation rate from approx 0.8 invocations/min to below 0.1 invocations when the mechanism triggering threshold is set over

30%. The shape of the curve is due to the fact that when the triggering threshold is large, the mechanism is barely invoked while when the threshold is too small, the execution rate is upper limited by the grace period setting of 75 seconds (i.e., 1/75 = 0.8/min). We also observe that the number of services does not largely impact the mechanism triggering rate.

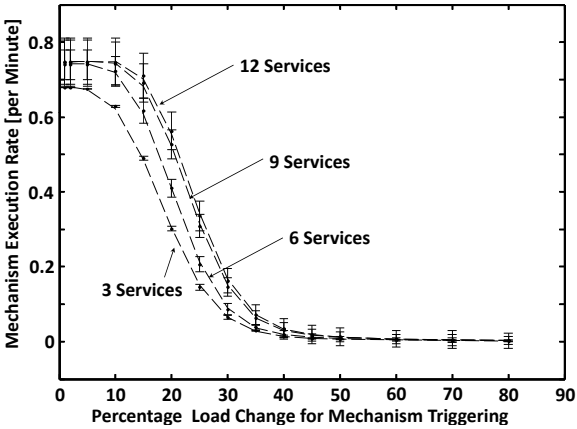


Figure 5.24: The effect of the mechanism triggering threshold (adapted from [145]).

Finally, we investigate the effect of the triggering threshold setting on the loss and the AAA system load. We use the same configuration as in Fig.5.22 with variable load and use the most granular threshold with 1% change in the load or session statistics as a reference. To compare to other threshold settings, we use the Root Mean Square Error (RMSE) for the load and the potential loss between the reference case (i.e., 1% threshold) and the threshold under consideration. The larger the RMSE, the worse the performance. As shown in Table 5.11, we observe that in our test case, the potential loss performance is affected significantly more than the system load by the triggering threshold. We also observe that the SCLP is the most sensitive scheme to the threshold setting while the APWC is the least sensitive. This is because the solution of the SCLP is not optimal and is more likely to fluctuate if not optimized frequently enough. On the other hand, the APWC tends to minimize the loss when the system is not overloaded and hence will not likely change the interim settings from the last optimal value. Table 5.12 provides a short comparison between the proposed accounting policies.

To sum up, we showed that the proposed mechanism maintains optimal service reporting intervals in dynamic environments which involve mobility, variation of the service load, tariff switching, and failovers. The results are encouraging as they proved that our mechanism is not only robust to various operational conditions but also showed that it is light weight and does not pose processing overhead on the system. The fact that the

Table 5.11: Root Mean Square Error for system load and normalized potential loss with reference to the 1% mechanism triggering threshold setting (adapted from [145]).

	Policy	Threshold Setting				
		5%	15%	25%	35%	45%
Loss	SCLP	1.2	1.8	6.2	9.1	13.2
	CLP	0.6	0.7	1.9	5.5	9.7
	APWC	0.4	0.5	1.0	1.6	1.6
Load	SCLP	0.2	0.3	0.6	1.0	1.2
	CLP	0.5	0.5	1.2	2.4	2.9
	APWC	0.5	0.5	0.9	1.6	1.6

Table 5.12: Summary and comparison between the accounting optimization policies (adapted from [145]).

		CLP	SCLP	APWC	Static
1.	**Loss guarantees**	Supported	Supported	Best Effort	Not Supported
2.	**Overload avoidance**	Medium/High	Medium/Low	Best	None
3.	**System requirements**	Changes to AAA + Optimization Solver	Changes to AAA	Changes to AAA + Optimization Solver	Already Supported
4.	**Execution time**	Medium [0.8-1.2]s	Very low [$<$ 20] ms	Medium/High [0.9-1.9]s	Lowest [$<$ 10]ms
5.	**Threshold sensitivity**	Medium Sensitive	Most Sensitive	Low Sensitivity	Not applicable
6.	**Complexity**	Polynomial Time [SQP]	One iteration	Polynomial Time [SQP]	No computation
7.	**Robustness and adaptability to changes**	Full	Full	Full	None

implementation scope of our mechanism is only limited to the AAA system supported by the promising results demonstrate its potential for commercial deployments.

5.6 Conclusions

In this section, we demonstrated the applicability of our planning framework to centralized and distributed AAA deployments in fixed and mobile networks under a wide range of protocol settings, mobility profiles, session statistics, and topological configu-

rations. We also evaluated the performance of our proposed optimization mechanisms relevant to mitigating authorization delay during handoffs as well as in optimizing accounting reliability in emerging multi-service cellular architectures. Relevant to AAA system planning, we attempted to answer the following primary questions that come to mind when planning network wide deployments for AAA systems,

- Under what conditions would statistical multiplexing design approaches used in fixed networks hold for designing AAA architectures in mobile networks ?
- How does mobility and roaming affect the design of AAA systems ?

In our results, we showed that statistical multiplexing approaches which scale the signaling rate only based on the session arrival rates and the session duration can be largely inaccurate in the context of mobile networks. We showed that this aspect is majorly attributed to mobility which is characterized by the AGW residence time and the mobility pattern between AGW regions. In our results, we observed that the fixed AAA signaling model can only be applied in low mobility scenarios where the AAA signaling rate distribution from each AGW matches that of the session arrival rates as one would expect. However, as this ratio approaches unity, ignoring mobility may lead to largely under-provisioning the AAA system. We have also shown that the signaling pertaining to multiple services does not necessarily match their arrival rates even if one sets their interim and reauthorization lifetime intervals proportionally to the service session duration. Again, this is a consequence of mobility between AGWs. However, we interestingly found that the fixed AAA model can approximate the signaling rate at the AAA system when context transfers are enabled between AGWs even in high mobility scenarios (error margin is below 15%). This is because context transfers between AGWs alleviate the mobility effects on the core network components including AAA systems. Our results also showed that in high mobility scenarios where the ratio of the interim interval or authorization lifetime settings to the mean residence time is relatively large, these messages may not be generated. However, setting such value low may result in very large signaling towards the AAA system as the users' mobility changes during the day. This issue is addressed by our accounting optimization framework. As such, from our results we concluded that using statistical multiplexing approaches along with the AAA signaling model for fixed networks is only applicable in special cases of context transfers or when mobility is low as indicated by the ratio of the mean session time to the residence time.

Relevant to mobility, we have shown that although the mobility pattern between AGWs may largely affect the observed signaling rate at AAA systems, it marginally or does not impact the AAA signaling rate in the following three cases: a) for centralized AAA systems serving home users if roaming is not considered b) when the roaming likelihoods from each AGW region to other networks is the same for centralized and distributed AAA architectures c) when mobility characterized by the mean session duration to the AGW residence time is low. For these conclusions to hold, the AGW residence times

must be i.i.d and one authentication protocol should be used by all AGWs. Under such conditions, the designer can assume any arbitrary mobility pattern and get accurate estimates of the signaling load. However, in cases where the mobility pattern is impacting such as for designing distributed AAA systems in general and for cases where signaling due to roaming users can not be neglected (e.g., when multiple MVNOs are served by an operator), our generalized AAA planning model offers designers with closed form analytical solutions which accommodate several protocol and topological scenarios. Such scenarios include the deployment of different authentication protocols at the AGWs, specific protocol optimizations for some AGWs (e.g., fast handoffs [31]), and the use of authentication delegation [9, 32] in roaming scenarios between gateway AAAs interacting with other networks and local AAA systems serving AGWs in the network under considerations. In an exemplary scenario, we have shown that authentication delegation can reduce the signaling load at the gateway AAA system by 40%.

Moreover, our planning model provides a granular view of the distribution of the AAA signaling load from each AGW for the possible four combinations of sessions (i.e., full, originating, terminating, and transit). This is useful in order to identify the bottleneck AGWs in the system and design distributed AAA systems. We have shown that the mobility pattern only affects the terminating and transit sessions while full and originating sessions follow the distribution of the users (i.e., the session arrivals). We have also shown that our model can be used to estimate the signaling load in centralized and distributed AAA systems irrespective of the supported protocols and even if the AGW residence times are not homogeneous. This is especially important for cases where mobile architectures are expected to support multiple wireless technologies such as WiMAX, EVDO, or LTE. In addition, we have shown that it can be applied for optimized authentication protocols such as EAP fast handoff (FH) schemes [31]. In our example, we showed that depending on the mobility pattern and residence times, the signaling reduction when EAP FH signaling is used can be in the vicinity of 10% to 20%. Finally, we have shown that the commonly used homogeneous residence time assumption in fact linearizes a non-linear effect on the signaling load in both roaming and non roaming scenarios. We have shown that such linear approximation can be acceptable for low mobility scenarios and depending on the mobility pattern can lead to large errors when mobility is high. To sum up, we have demonstrated the applicability of our AAA planning framework and its wide range of applications to both centralized and distributed AAA system deployments including different authentication protocols, user distributions, mobility profiles, roaming, and non-homogeneous AGW residence times. Further research avenues for our planning work include the derivation of higher order moments for the AAA signaling rate, relaxing the exponential session time assumption, and developing queuing models for AAA messages to comprehensively analyze the authorization delay for real time services.

With respect to AAA performance optimization, we focused on the scalability and performance of the proposed authentication delay mitigation and accounting reliability optimization schemes after demonstrating their basic operation in Chapter 4. In our discussion, we addressed the following issues,

- Would the proposed handoff QoS signaling mechanism still scale similarly to the original scheme as function of session arrival rates and network size ?
- How do our accounting optimization policies behave in multi-service environments under varying load and mobility conditions ?

Relevant to the scalability of our authentication mechanism, we investigated the mechanism's scalability in an exemplary EVDO network as function of the session arrival rate, number of cells per RNC, cell radii, users' concentration in the handoff zone, and mobility pattern. Using numerous simulations, we have shown that our proactive mechanism scales similarly to the original authentication mechanism and that it does not impose a significant load on the serving cellular network. In our evaluations, we have shown that although the AGW handoff rate increases significantly as the cell size is decreased, the proposed mechanism's signaling scales similarly to the original procedure. Based on two exemplary movement patterns with directed movement where movers do not change their direction within spans of 8km and another random pattern where direction is changed every 200m, we established bounds on the signaling rate due to our mechanism. This is because directed movers barely change direction and hence each proactive signaling operation is most likely followed by a handoff while for random movers this is not the case. Our results showed that proactive signaling is triggered at approximately double the handoff rate for random movers and approximately at the handoff rate for directed users. Another interesting finding that simulations revealed is that our mechanism scales similarly to the standard mechanism as the number of cells (x) per RNC is changed and with a complexity of $O(\frac{1}{\sqrt{x}})$ which matches known results for the number of handoffs in square cellular arrangements [144]. The results also showed that even if a relatively large user concentration (i.e., 25%) exists in the border region between RNCs, the proposed mechanism still scales similarly to the standard mechanism.

For completeness, we have also investigated the signaling on relevant network interfaces including AAA-RNC, AAA-AGW, AAA-PCRF, PCRF-AS, and PCRF-AGW. We have observed that for highly directional movers, the signaling load of the proactive and the standard mechanisms is almost the same on all interfaces except on the AAA-RNC interface due to the handoff imminent (HI) notifications. On the other hand, for random movers and due to the large likelihood of false alarms, the signaling rate for the proactive mechanism is approximately three-folds that of the standard mechanism on the AAA-RNC interface due to the more frequent transmission of the HI messages and is almost two-folds on rest of the interfaces. In real deployments, a spectrum of movement patterns is observed and hence the observed signaling rates on the corresponding network interfaces will be in between that of random and directional movers. Future work avenues to enhance our mechanism is to investigate its integration with evolving standards such as the IEEE 802.21 framework [132] and to identify cases where inter-operability between RADIUS and Diameter signaling can pose operational issues.

Relevant to our accounting optimization policies, we have demonstrated robustness and stability of the proposed polices in controlling the potential loss and the AAA signaling

load under various scenarios of service durations and costs, tariff switching, mobility, roaming, and AGW failovers. We have shown that the Constrained Loss Policy (CLP) offers more optimal load performance than the Simplified Constrained Loss Policy (SCLP) while both maintain the same potential loss target. The loss from the APWC mechanism in our cases usually minimized the interim interval setting for most services that are not flat rate (i.e., posing zero potential loss) unless the AAA system load is high. The results clearly showed that our policies allow much better control of the potential loss relative to static policies (i.e., fixed interim interval settings) and are more resilient to changes in session statistics as they manage to either minimize the loss (i.e., in the APWC case) or maintain a constant loss target (i.e., in the CLP and SCLP cases) in fixed and mobile networks. In addition, in an exemplary scenario of AGW fail-over, we have shown how our policies are all able to handle failover scenarios and can minimize the potential loss by four times on the cost of only 10% additional signaling load compared to a static policy that sets the interim interval equal to the mean session duration. We have also illustrated the operation of our mechanism in roaming scenarios over an AAA proxy chain by having only one AAA system running the optimization policies over a separate pre-reserved signaling capacity among all AAAs in the chain.

We also investigated the processing load of each of the proposed policies relevant to the number of supported services and the frequency of the mechanism's invocations. We observed that the number of services does not largely impact the mechanism's triggering rate. The results revealed that the SCLP requires time below 1% of the other policies regardless of the AAA system loading while the APWC policy was the most affected by the AAA system load due to the non-linearity of its objective function. On the other hand, the results showed that the SCLP requires the most number of optimizations to maintain the balance between the potential loss and the signaling load while the APWC scheme requires the least number of optimizations. This is because the solution of the SCLP is not optimal and is more likely to fluctuate if not optimized frequently enough. Furthermore, the APWC tends to minimize the loss when the system is not overloaded and hence will not likely change the interim settings from the last optimal value. In conclusion, the results are encouraging as they proved that our mechanism is not only robust to various operational conditions but also showed that it is light weight and does not pose appreciable processing overhead on the system. The fact that the implementation scope of our mechanism is only limited to the AAA system supported by the promising results demonstrate its potential for commercial deployments. Future work includes the implementation of the mechanism using open source AAA packages such as Free RADIUS or Open Diameter and validating its performance with real captures of accounting traffic. Moreover, our mechanism can be extended to support unified billing architectures combining both prepaid and postpaid mechanisms as well as integration with dynamic pricing tools in the business support system [54].

To sum up, in this section we demonstrated the scope and the applicability of our AAA system planning models to cover centralized and distributed AAA system deployments under a wide range of design variables including session durations, mobility profiles, and protocol settings. We have also illustrated the robustness of the proposed optimiza-

tion mechanisms for authentication delay and accounting reliability in multi-service networks for different network sizes, mobility, failovers, and roaming. As such, we believe that our planning models can efficiently guide the design and testing of AAA systems in mobile systems and help avoid large over-provisioning and arbitrary designs. We also think that the results for the proposed optimization schemes carry promising potential towards integration within commercial AAA products for multi-service mobile networks.

Chapter 6 Conclusions and Future Work

In this thesis, we studied a set of problems relevant to the design of Authentication, Authorization, and Accounting (AAA) systems in mobile telecommunications networks including system planning, protocol optimizations, and new extensions and applications. With respect to AAA system planning, we have developed closed form analytical models for the signaling rate towards the AAA system from access gateways under a wide range of protocol settings, network topology, and mobility parameters. We then proposed an optimization framework which mitigates the AAA authentication delay and enhances the AAA accounting reliability in multi-service environments. We have also proposed extensions for AAA frameworks in the areas of cellular backhaul over Wireless Mesh Networks (WMN) and in layer 2 optical communications between different carriers. Using several case studies, we demonstrated the applicability of our AAA planning models and extensions for a range of parameters and illustrated the scale of our optimization mechanisms as function of session and mobility statistics as well as network topology.

Our work on AAA system planning offers the first foundational framework which answers the pivotal question of how the AAA signaling rate relates to protocol parameters, mobility, and network topology. Since the size of the AAA systems in the network is primarily determined by the signaling rates from the serving IP Access Gateways (AGWs), we proposed analytical planning framework for the signaling rate which considers relevant AAA protocol parameters such as the accounting interim and authorization lifetime intervals for centralized and distributed AAA mobile network deployments. In our development, we started by analyzing the case where AAA systems are deployed in fixed network scenarios where mobility plays no role. In this case, only the session arrival rates and AAA protocol settings such as the accounting interim intervals and authorization lifetimes are relevant. Under these assumptions, our analysis revealed that the signaling rate is a function of the discretized Complementary Commutative Distribution Function (CCDF) of the session duration at the interim interval and authorization lifetime. However, we observed that assuming no mobility between AGW regions can easily lead to under provisioned systems due to the presence of users residing on the border regions between AGWs. In fact, a concentration as low as 10% of such users can lead to 60-100% under provisioning depending on the session statistics. As emerging networks are expected to become flatter (i.e., reduced hierarchy), such effects are expected to grow considerably.

To this end, we generalized our analysis for mobile networks by utilizing concepts of holding and residence time from cellular performance theory. For all AAA deployments, we related the number of interim and reauthentication messages to the time duration an AGW serves a session (i.e., the AGW holding time). Depending on the mobility of the users during their sessions, we identified four possible types of AGW holding times. We showed that the number of reauthentication and interim messages directly depends on the discretized CCDF of the duration of these holding times at the interim interval and authorization lifetime. An interesting outcome of our analysis besides the determination of the AAA signaling load is the identification of a potential operational region which operators can use to select their interim intervals and authorization lifetime settings. In this operational region which extends between one half to full session duration, the signaling rate changes almost linearly as function of the mean session duration. Finally, we concluded our discussion on AAA system planning by generalizing our analysis to handle multiple AAA systems and incorporate additional design variables including the possibility of having different authentication protocols, the mobility pattern between AGW regions and cells, and the likelihood to roam to other partner networks. Our planning model utilizes concepts of Markovian mobility models and offers closed form solutions for the signaling rate in the network. Put simply, the design implications of our planning framework are two folds: first, for a system planner, bounds on the AAA signaling load can be quickly estimated for different services based on the expected or measured service usage and mobility statistics. Second, systems quality assurance engineers can accurately configure their load generating tools based on service statistics and protocol settings and hence avoid time consuming trial and error in generating the desired signaling rates.

With respect to AAA system optimization, we identified two issues in the context of multiservice networks: the mitigation of the authentication delay and the optimal choice for accounting interim intervals for the reliability of accounting records. Relevant to the first, we focused on the reduction of QoS signaling delay for third party services using proactive signaling facilitated by the AAA framework. This is because QoS signaling can result in undesirably variable and prolonged signaling delays (of the order of seconds) upon AGW handoffs as the policy system, which uses AAA protocols, may contact one or more application servers for QoS authorization. To address this issue, we proposed a proactive signaling mechanism in the application layer that conveys authorization delay constraints from the service layer to the radio layer and thus mitigates the effects of variable signaling delays. The proposed mechanism exploits the already established mechanisms of authentication and authorization signaling via standardized interfaces and protocols and goes inline with the proactive signaling mechanisms in the IEEE 802.21 framework. Our evaluation showed that the proposed mechanism is able to minimize the authorization delay variations due to round trip delay between the policy system and application servers and with other policy systems when roaming, and due to variations in the application servers' load.

Relevant to accounting optimization, we addressed the optimal setting of the accounting interim intervals in multi-service environments such that the incurred capital losses is

minimized if the serving Network Access Server (NAS), which can be an AGW, fails. This is because although short accounting interim intervals minimize the potential loss, they are likely to result in undesirably high signaling load especially when different service flows are supported. We showed that this problem is non-trivial as it primarily involves considering cost and statistical properties of multitudes of services with different mobility profiles. We proposed a dynamic policy based optimization mechanism which is based on AAA standards and does not require changes to the NASes in the network. We demonstrated that the proposed mechanism limits the potential loss in the event of Network Access Server (NAS) failure without excessively generating unnecessary usage reports.

In addition to AAA optimization mechanisms, we have proposed the introduction of AAA systems to novel areas such as cellular backhaul and layer 2 optical networking. Relevant to the first, we argued that WMN operators can become players in the cellular backhaul industry and hence there is a clear need for AAA solutions in that area. Therefore, we proposed the first billing architecture for cellular backhaul applications over WMNs in conjunction with a simple threshold based bandwidth management algorithm for backhaul connections. In our design, adding or releasing backhaul bandwidth chunks generates billing updates and hence needs to be carefully investigated. Hence, we evaluated a relatively unstable reservation mechanism and established the corresponding upper bounds on the billing signaling performance over WMN backhauls. We found that even a poor reservation scheme can be accommodated by current commercial hardware. On the other end of the spectrum, we proposed the incorporation of AAA signaling for inter-carrier path provisioning in optical networks after the success of the Path Computation Element (PCE) framework for multi-carrier environments. This scenario is particularly interesting as the data path may traverse more than two networks; which suggests that all the participating networks authorize data path provisioning operations and implement their metering functionalities. The proposed mechanism addresses such challenges and was specifically designed to facilitate secure exchange of path computation signaling among domains, associate path setup with the paths computed by the PCE, while enabling sharing of accounting information between carriers. The analysis and results showed that the signaling mechanism is light weight and may be integrated within the PCE platform, which demonstrates potential for commercial deployments.

Finally, we concluded this thesis with multiple study cases that demonstrate the applicability of our AAA planning models and the scalability of our optimization schemes. For different mobility profiles, we have shown that the fixed AAA signaling model can only be applied in low mobility scenarios when the ratio of the session to residence time is below unity. Furthermore, for medium and highly mobile users, we have observed that the signaling pertaining to multiple services does not necessarily match their arrival rates even if one sets their interim and reauthorization lifetime intervals proportionally to the service session duration. However, we interestingly found that fixed AAA model can approximate the signaling rate at the AAA system when context transfers are enabled between AGWs even for relatively high mobility users (error margin is below 15%). Another interesting finding that our analysis revealed is that irrespective of the

mobility profile of the users, any arbitrary mobility pattern between AGW regions results in the same AAA signaling rate in cases of centralized AAA system deployments with insignificant roaming traffic or when roaming likelihoods to other networks from each AGW region are similar. These results hold if the AGW residence times are i.i.d and one authentication protocol is used.

Using exemplary network configurations served by multiple AAA systems and implementing different authentication protocols, we demonstrated the ability of our planning model to capture the signaling rate at each AAA system and from each AGW in the network. We have also demonstrated the capability of the model to obtain the signaling rate for home and roaming users which result in proxy AAA requests from the AAA infrastructure towards their networks. Moreover, we have shown the capability of the model to give estimates for special optimizations such as authentication delegation between AAAs for roaming users and fast handoffs which reduce the signaling load and delay during handoff events. Finally, we have also shown that using the same residence time statistics for all regions in the network linearizes an inherently non-linear effect on the signaling load. We have shown that such linear approximation can be acceptable for low mobility scenarios and depending on the mobility pattern can lead to large errors when mobility is high.

We also investigated the scalability of our authentication optimization mechanism in an exemplary EVDO network as function of the session arrival rate, number of cells, cell radii, users' concentration in the handoff zone, and mobility pattern. Using numerous simulations, we have shown that our proactive mechanism scales similarly to a system that does not implement our proactive scheme. By evaluating the signaling due to directed and random movers, we established bounds on the signaling rate due to our mechanism in all relevant network interfaces. Our results showed that proactive signaling is triggered at approximately double the handoff rate for random movers and almost equal to the handoff rate for directed movers since they barely change their direction. Other patterns result in a rate that falls within these bounds. We also showed that our mechanism scales similarly to the original system and proportionally to the inverse of the square root of the number of cells in a RNC region and is not impacted even if a relatively large user concentration exists in border regions between AGWs.

Moreover, we have investigated the scalability of our accounting optimization mechanism and demonstrated its robustness and stability in controlling the potential loss and the AAA signaling load. We examined its behavior under various scenarios of service durations and costs, tariff switching, mobility, roaming, and AGW failovers. The results clearly showed that our optimization policies allow much better control of the potential loss compared to the static policies (i.e., fixed interim interval settings). Our optimization policies were more resilient to changes in session statistics as they manage to either minimize the loss (i.e., in the Adaptive Policy with Weight Control (APWC) case) or maintain a constant loss target (i.e., in the Constrained Loss Policy (CLP) and Simplified CLP (SCLP) cases) in fixed and mobile networks. We have also illustrated how our policies are able to handle failover scenarios and can minimize the potential loss by four

times for an extra signaling load of less than 10% compared to static policies with reasonable interim interval settings. The results also showed that the number of services does not largely impact the mechanism's invocation rate and revealed that the SCLP requires time below 1% of the other policies regardless of the AAA system loading.

Although we endeavored to address various design and optimization aspects of AAA systems in mobile networks, future research is still required at the system architecture, mathematical, and implementation levels. Relevant to the first, further work is still needed to address emerging trends of converged billing systems where prepaid and postpaid accounting records are processed by a common Business Support System (BSS) in real time. The challenges in this direction include the lack of a solid understanding of prepaid systems due to the unavailability of statistics relevant to the users' consumption patterns, their quotas, and their churn rates as well as the fact that most of the BSS is heavily based on proprietary solutions. The real time nature for the operation of converged billing systems makes the knowledge of the accounting signaling rate vital as it not only determines the size of the AAA system but also the load on numerous other BSS components such as rating, balance management, and fraud management systems. This also entails enhancements to our proposed accounting reliability mechanism to handle aspects of prepaid systems including the signaling rate between the AAA system and the prepaid server, and effects related to the users' balance. Eventually, we propose that our mechanism is extended to optimize accounting settings not only based on the AAA system but also by including systems within the converged BSS such as the rating and balance management systems. Finally, with the emergence of cloud computing paradigms, billing systems can be offered as a service (a.k.a., BaaS) which brings further challenges due to virtualization relevant to the operation (e.g., server assignment and load balancing) and billing for using the cloud (i.e., how users of the cloud are charged). As such planning models and optimization should be further evolved to assist load balancing as well as consider the possibility of task migration within the cloud systems in addition to accounting parameters in order to control the potential loss.

Second, at the mathematical level, further work is still needed to generalize our analysis for the signaling rate for higher order statistics as well as the consideration of generalized session distributions. The first entails complexity in the consideration of correlation within the signaling stream which depends on the accounting interim interval and authorization lifetime settings as well as the higher order statistics of the number of handoffs for users with high mobility profiles. The generalized session distribution is also complex as it entails the derivation of the holding times as functions of the handoff history. Up to our knowledge, this problem was not solved yet at the fundamental level in cellular performance studies. Moreover, due to the foreseen real-time nature for emerging systems, research is needed to determine the most suitable queuing models for AAA systems and eventually for converged billing paradigms based on the session arrival process. The queuing models will determine the delay metrics of the billing system. In addition, analytical models are needed to determine the load on prepaid systems in time and volume based charging models using general statistical fits of the users' account balances. Finally, sensitivity analysis is also needed to determine the dependence

of the planning models on each design variable. This is pivotal as such analysis offers insights on the required accuracy level for the variable estimates and statistical fits.

Third, from an implementation perspective, future work is also needed to develop our proposed mechanisms for optimization using freely available AAA open source packages such as Free RADIUS and Open Diameter. This includes avenues for different deployment options for the functional entities of our optimization mechanisms. For instance, in our accounting method, the optimization solver can reside in one server while the rest of the functional entities may be deployed within each AAA system independently. For our proactive authorization mechanism, experimental test beds where the wireless signal is attenuated according to a simulated movement track in a prototype cellular area can be used to test the mechanism's robustness. Future work is also needed to integrate our proposed mechanisms within ongoing standardization efforts such as the IEEE 802.21 and Diameter standards. Finally, from a testing and validation perspective, further work is required to implement open source AAA traffic generation tools based on realistic data measurements or using simulation models. This allows testing the performance of the AAA system and the proposed mechanisms and models in relation with the deployed users' databases (e.g., SQL or LDAP based databases) and systems within the BSS. The load generation tools shall rely on measurements of the users' mobility profiles and service duration/volume statistics.

To sum up, in this thesis, we have developed analytical AAA planning models, designed optimization schemes in multi-service environments, and proposed new applications for AAA protocols in cellular backhaul and optical communications applications. The results demonstrated the flexibility, the robustness, and the applicability of our planning models to a spectrum of design variables including protocol settings, mobility, session statistics, and network topology. Future research avenues for this thesis fall in the areas of designing converged billing systems, developing queuing models for the AAA signaling load, and validating the work using real accounting records in representative test-beds.

Appendices

Appendix A Proofs

A.1 Proofs from Chapter 3

A.1.1 Proof of (3.15)

To do so, let J denote the random number of interims in an arbitrary duration, H. Then, we can write $J = \lfloor \frac{H}{\Delta_T} \rfloor$. As shown in Figure A.1, the PDF of J is then obtained by integrating over all the area between $[j\Delta_T, (j+1)\Delta_T]$ as [102],

$$\begin{aligned} f_J(j) &= \int_{j\Delta_T}^{(j+1)\Delta_T} f_H(h)\,dh = F_H((j+1)\Delta_T) - F_H(j\Delta_T) \\ &= \bar{F}_H(j\Delta_T) - \bar{F}_H((j+1)\Delta_T) \end{aligned} \tag{A.1}$$

where $\bar{F}_H(h)$ denotes the CCDF of the duration h and is given as $\bar{F}_H(h) = 1 - F_H(h) = 1 - \int_{-\infty}^{h} f_H(x)\,dx$. Using (A.1), the mean number of interims or reauthentications (i.e., $\varphi(\Delta_T)$ and $\varphi(\Delta_M)$ respectively) during an arbitrary duration, H, is given by calculating the expectation of the floor of H by the corresponding period (e.g, $E[\lfloor \frac{H}{\Delta_T} \rfloor]$) as,

$$\begin{aligned} \varphi(\Delta_T) = E[\lfloor \frac{H}{\Delta_T} \rfloor] &= \sum_{j=0}^{\infty} j f_J(j) = \sum_{j=0}^{\infty} j \left[\bar{F}_H(j\Delta_T) - \bar{F}_H((j+1)\Delta_T) \right] \\ &= \sum_{j=1}^{\infty} j\bar{F}_H(j\Delta_T) - \sum_{j=1}^{\infty} (j-1)\bar{F}_H(j\Delta_T) = \sum_{j=1}^{\infty} \bar{F}_H(j\Delta_T) \end{aligned} \tag{A.2}$$

A.1.2 Evaluating the Holding Times

Before we start evaluating the specific CDF of each distribution, we show the method of calculating two important probabilities: $(X < Y)$ and $(X < z_0 \mid X < Y)$ where X and Y are non-negative random variables and z_0 is a positive constant. From probability

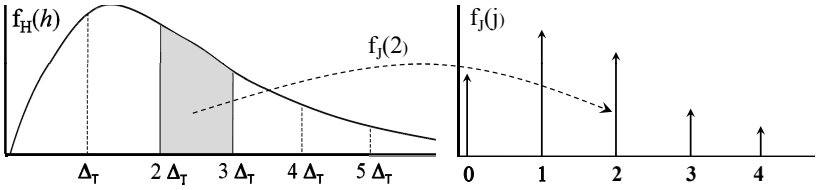


Figure A.1: General procedure for obtaining the PDF of the floor of a random variable, X

theory, we know that the probability $(X < Y)$ is evaluated as follows,

$$\Pr\{X < Y\} = \int_{y=0}^{\infty} \int_{x=0}^{y} f_X(x) f_Y(y) \, dxdy \tag{A.3}$$

A closed form solution for (A.3) can be obtained using the residue theorem as discussed in [17]. When X follows the exponential distribution, a simple result is obtained as,

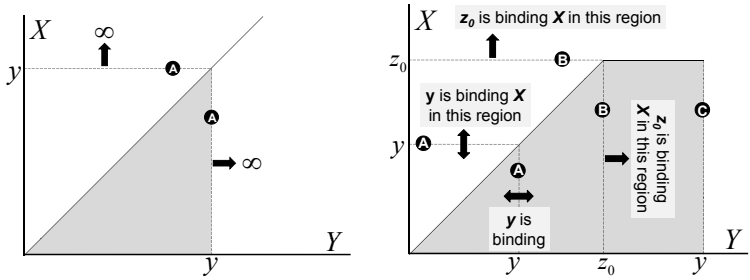
$$\begin{aligned} \Pr\{X < Y\} &= \frac{1}{E[X]} \int_{x=0}^{\infty} \bar{F}_Y(x) \, e^{-\frac{x}{E[X]}} \, dx = \frac{1}{E[X]} \mathscr{L}\left\{ \frac{1 - f_Y^*(\hat{s})}{\hat{s}} \mid_{\hat{s}=\frac{1}{E[X]}} \right\} \\ &= 1 - f_Y^*(\frac{1}{E[X]}) \end{aligned} \tag{A.4}$$

The integration region in (A.3) is shown in Figure A.2(a) and always pertains to the shaded triangular area for all possible values of X and Y. On the other hand, to evaluate the conditional probability that $(X < z_0 \mid X < Y)$, we use Bayes' theorem as follows,

$$\Pr\{X < z_0 \mid X < Y\} = \frac{\Pr\{X < z_0 \cap X < Y\}}{\Pr\{X < Y\}} \tag{A.5}$$

The denominator of (A.5) can be evaluated using (A.3). To evaluate the joint probability $\Pr\{X < z_0 \cap X < Y\}$ in the numerator, we need to identify the integration region after introducing the restriction that $X < z_0$. Notice that the restriction is only on X not on Y and hence the integration over Y should be carried out over all possible values of Y. The integration region is depicted in Figure A.2(b)and we have two cases as follows,

- The integration variable y is less than z_0 (i.e., $0 \leq y < z_0$): In this case, X is limited by y and hence we have a triangular area as in Figure A.2(a).
- The integration variable y is greater than or equals z_0 (i.e., $y \geq z_0$): In this case, we have a rectangular area as X is always limited to z_0 as y approaches ∞.



(a) Region: $(X < Y)$ (b) Region: $(X < z_0 \cap X < Y)$

Figure A.2: The integration region for (a) $X < Y$ and (b) $X < z_0 \cap X < Y$.

Thus, the joint probability $\Pr\{X < z_0 \cap X < Y\}$ is evaluated as follows,

$$\begin{aligned}\Pr\{X < z_0 \cap X < Y\} &= \int_{y=0}^{\infty}\int_{x=0}^{min(y,z_0)} f_X(x) f_Y(y)\, dx\, dy \qquad (A.6)\\ &= \int_{y=0}^{z_0}\int_{x=0}^{y} f_X(x) f_Y(y)\, dx\, dy + \int_{y=z_0}^{\infty}\int_{x=0}^{h_0} f_X(x) f_Y(y)\, dx\, dy\\ &= \int_{y=0}^{z_0} F_X(y) f_y(y)\, dy + F_X(z_0)\int_{y=z_0}^{\infty} f_Y(y)\, dy\\ &= \int_{y=0}^{z_0} F_X(y) f_y(y)\, dy + F_X(z_0)\bar{F}_Y(z_0)\end{aligned}$$

Substituting (A.6) and (A.3) into (A.5), we have,

$$\Pr\{X < z_0 \mid X < Y\} = \frac{\int_{y=0}^{z_0} F_X(y) f_y(y)\, dy + F_X(z_0)\bar{F}_Y(z_0)}{\int_{y=0}^{\infty}\int_{x=0}^{y} f_X(x) f_Y(y)\, dxdy} \qquad (A.7)$$

Finally, the PDF of the CDF in (A.7) is given as,

$$\frac{d\Pr\{X < z_0 \mid X < Y\}}{dz_0} = \frac{1}{\Pr\{X < Y\}} f_x(z_0)\bar{F}_y(z_0) \qquad (A.8)$$

A.1.3 Proof of (3.28)

To prove (3.28), we show that:

$$F_{H_O}(h) = Pr\left(S \le h \mid \tilde{R} \le S\right) \equiv F_{H_T}(h) = \Pr\{S \le h \mid S \le R\}$$

This can be achieved by showing the PDFs are equivalent (i.e., $f_{H_O}(h) = f_{H_T}(h)$). Since $f_{\bar{R}}(\bar{r}) = \frac{\bar{F}_R(r)}{E_r}$, then using (A.8) we have,

$$f_{H_O}(h) = \frac{f_{\bar{R}}(h)\bar{F}_S(h)}{\Pr\{\bar{R} < S\}} = \frac{\frac{E_S}{E_R}\bar{F}_R(h)f_S(h)}{\frac{E_S}{E_R}\left(1 - f_R^*\left(\frac{1}{E[X]}\right)\right)} = \frac{\bar{F}_R(h)f_S(h)}{1 - f_R^*\left(\frac{1}{E[X]}\right)}$$

where $f_R^*(\hat{s})$ denotes the Laplace transform of $f_R(r)$ as $\mathscr{L}\{f_R(r)\}$. Using a similar approach to evaluate $f_{H_T}(h)$, we have

$$f_{H_T}(h) = \frac{f_S(h)\bar{F}_R(h)}{\Pr\{S < R\}} = \frac{\bar{F}_R(h)f_S(h)}{\left(1 - f_R^*\left(\frac{1}{E[X]}\right)\right)} = \frac{\bar{F}_R(h)f_S(h)}{1 - f_R^*\left(\frac{1}{E[X]}\right)}$$

Since $f_{H_T}(h) = f_{H_O}(h)$, then $H_O \equiv H_T$.

A.1.4 Proof of (3.93) and (3.94)

In the following proposition we avoid the inversion operation by studying (3.88) as,

Proposition A.1.1. *Let the matrix* $\mathbf{Q}^{(\mathbf{x})}$ *of* (3.88) *be truncated to an arbitrary finite number of handoffs k less than K (i.e.* $\mathbf{Q}^{(\mathbf{x})}$ *contains sub-matrices up to* $\mathbf{D_K}$*), then the matrix* $\mathbf{M}_\mathbf{z}^{(\mathbf{x})} = \left(\mathbf{e} - \mathbf{Q}^{(\mathbf{x})}\right)^{-1}$ *has the structure*

$$\mathbf{M} = \begin{pmatrix} \mathbf{I} & \mathbf{D_0} & \mathbf{D_0D_1} & \mathbf{D_0D_1D_2} & \mathbf{D_0D_1D_2D_3} & \cdots \\ 0 & \mathbf{I} & \mathbf{D_1} & \mathbf{D_1D_2} & \mathbf{D_1D_2D_3} & \cdots \\ 0 & 0 & \mathbf{I} & \mathbf{D_2} & \mathbf{D_2D_3} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \tag{A.9}$$

Equation (A.9) *is simple to obtain by solving the linear system* $\mathbf{M}_\mathbf{z}^{(\mathbf{x})}\left(\mathbf{e} - \mathbf{Q}^{(\mathbf{x})}\right) = \mathbf{e}$ *and applying principles of induction.*

Using (3.91) and (A.9), the mean number of handoffs before leaving the network $E\{K_x\}$ in (3.91) is given as,

$$E\{K_x\} = \mathbf{P}_\mathbf{I}^{(\mathbf{x})}\mathbf{M}_\mathbf{z}^{(\mathbf{x})}\mathbf{o}^T - 1 = \mathbf{F}^{(\mathbf{x})}\left(\mathbf{e} + \sum_{k=0}^{\kappa}\prod_{j=0}^{k}\mathbf{D_j}\right)\mathbf{o} - 1 \tag{A.10}$$

The (-1) cancels out in (A.10) since $\mathbf{F}^{(\mathbf{x})}\mathbf{eo} = 1$. Let us define $(\mathbf{Q_M})^0 = \mathbf{e}$, then utilizing

(3.89) we get,

$$E\{K_x\} = \mathbf{F}^{(\mathbf{x})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})}\left(\sum_{k=0}^{\kappa}(\mathbf{Q_M})^k\prod_{j=0}^{k}\gamma_{(j+1)}^{x}\right)\mathbf{o}$$
$$= \mathbf{F}^{(\mathbf{x})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})}\left(\sum_{k=0}^{\kappa}G^{(x)}(k+1)\left(\mathbf{Q}_{\mathbf{M}}^{(\mathbf{x})}\right)^k\right)\mathbf{o} \tag{A.11}$$

where we used the observation that $\gamma_k^x = \frac{G^{(x)}(k)}{G^{(x)}(k-1)}$ and $G^{(x)}(0) = 1$, (i.e., $\prod_{j=0}^{k}\gamma_{(j+1)}^{x} = \prod_{j=0}^{k}\frac{G^{(x)}(j+1)}{G^{(x)}(j)} = G^{(x)}(k+1)$).

Using the complex integral representation for $G^{(x)}(k)$ in (3.86) and letting the limit $\kappa \to \infty$, we get,

$$E\{K_x\} = \mathbf{F}^{(\mathbf{x})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})}\left(\sum_{k=0}^{\kappa}\frac{\mathbf{Q}_{\mathbf{M}}^{(\mathbf{x})k}}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_{R_1}^*(\hat{s})f_S^*(-\hat{s})}{\hat{s}}\left(f_R^*(\hat{s})\right)^k d\hat{s}\right)\mathbf{o}$$
$$= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_{R_1}^*(\hat{s})\mathbf{F}^{(\mathbf{x})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})}f_S^*(-\hat{s})}{\hat{s}}\sum_{k=0}^{\kappa}\left(f_R^*(\hat{s})\mathbf{Q}_{\mathbf{M}}^{(\mathbf{x})}\right)^k\mathbf{o}d\hat{s}$$
$$= \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\frac{f_{R_1}^*(\hat{s})\mathbf{F}^{(\mathbf{x})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})}f_S^*(-\hat{s})}{\hat{s}}\mathbf{M}_{\mathbf{R}}^{(\mathbf{x})}(\hat{s})\mathbf{o}d\hat{s}$$

where we have defined the matrix $\mathbf{M}_{\mathbf{R}}^{(\mathbf{x})}$ as,

$$\mathbf{M}_{\mathbf{R}}^{(\mathbf{x})}(\hat{s}) = \left(\mathbf{e} - f_R^*(\hat{s})\mathbf{Q}_{\mathbf{M}}^{(\mathbf{x})}\right)^{-1} \tag{A.12}$$

and that R_1 denotes the AGW residence time in the first AGW serving the session. Thus, $f_{R_1}^*(\hat{s}) = f_{\bar{R}}(\hat{s})$ for on-net sessions and $f_{R_1}^*(\hat{s}) = f_R(\hat{s})$ for off-net sessions.

Using the residue theorem, we obtain the final closed form solution for the mean number of handoffs inside the network, $E\{K_x\}$ as,

$$E\{K_x\} = -\sum_{s_P\in\Xi_{\hat{S}_-}}\operatorname*{Res}_{\hat{s}=s_p}\frac{f_{R_1}^*(\hat{s})\mathbf{F}^{(\mathbf{x})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})}f_S^*(-\hat{s})}{\hat{s}}\mathbf{M}_{\mathbf{R}}^{(\mathbf{x})}(\hat{s})\mathbf{o} \tag{A.13}$$

Similarly, using (A.9) and assuming that $\mathbf{A_{K+1}} = \mathbf{0}$, the roaming probability, β_x, in

(3.92) is given as,

$$\beta_x = \mathbf{P}_{\mathbf{I}}^{(\mathbf{x})}\mathbf{M}_{\mathbf{z}}^{(\mathbf{x})}\mathbf{A} = \mathbf{F}^{(\mathbf{x})}\mathbf{A_0} + \mathbf{F}^{(\mathbf{x})}\left(\sum_{k=1}^{\kappa}\prod_{j=0}^{k-1}\mathbf{D_j}\mathbf{A_k}\right) \tag{A.14}$$

$$= G^{(x)}(1)\,\mathbf{F}^{(\mathbf{x})}\mathbf{A_{MI}} + \mathbf{F}^{(\mathbf{x})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})}\left(\sum_{k=1}^{\kappa}G^{(x)}(k+1)\left(\mathbf{Q}_{\mathbf{M}}^{(\mathbf{x})}\right)^{k-1}\right)\mathbf{A_M}$$

Using the complex integral presentation, the second term in (A.14) denoted as $\hat{\beta}_x$ is given as,

$$\hat{\beta}_x = \frac{\mathbf{F}^{(\mathbf{x})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})}}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty}\left[\frac{f_{R_1}^*(\hat{s})\,f_R^*(\hat{s})\,f_S^*(-\hat{s})}{\hat{s}}\sum_{k=1}^{\kappa}\left(f_R^*(\hat{s})\,\mathbf{Q_M}\right)^{k-1}\mathbf{A_M}\right]d\hat{s}$$

Taking the limit $\kappa\to\infty$ and applying the residue theorem, we get a closed form expression for the roaming probability β_x as,

$$\beta_x = G^{(x)}(1)\,\mathbf{F}^{(\mathbf{x})}\mathbf{A_{MI}} - \sum_{s_P\in\Xi_{\hat{S}_-}}\operatorname*{Res}_{\hat{s}=s_p}\frac{f_{R_1}^*(\hat{s})\,\mathbf{F}^{(\mathbf{x})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{x})}f_S^*(-\hat{s})}{\hat{s}}\mathbf{M}_{\mathbf{R}}^{(\mathbf{x})}(\hat{s})\,f_R^*(\hat{s})\,\mathbf{A_M} \tag{A.15}$$

A.1.5 Example of using (3.93)

As an application we consider the case, where the residence time is generally distributed and the session time follows a hyper-Erlang distribution with Laplace transform,

$$f_S^*(s) = \sum_{j=1}^{J}\alpha_j\left(\frac{\mu_j}{s+\mu_j}\right)^{m_j}$$

The poles of $f_S^*(-s)$ are located at $s_{Pj} = \mu_j > 0$ and hence satisfies the residue theorem. Thus, using (3.93) for on-net traffic we get,

$$E\{K_\Omega\} = -\sum_{j=1}^{J}\frac{\alpha_j(-\mu_j)^{m_j}\,\mathbf{F}^{(\mathbf{\Omega})}\mathbf{Q}_{\mathbf{MI}}^{(\mathbf{\Omega})}}{(m_j-1)!\,E\{R\}}\lim_{\hat{s}\to\mu_j}\frac{d^{m_j-1}}{d\hat{s}^{m_j-1}}\mathbf{V_N}(\hat{s})\,\mathbf{o}$$

where we have defined the term $\mathbf{V_N}$ as,

$$\mathbf{V_N}(\hat{s}) = \frac{1-f_R^*(\hat{s})}{\hat{s}^2}\mathbf{M}_{\mathbf{R}}^{(\mathbf{\Omega})}(\hat{s}) = \frac{1-f_R^*(\hat{s})}{\hat{s}^2}\left(\mathbf{e}-f_R^*(\hat{s})\,\mathbf{Q_M}\right)^{-1}$$

For the practically interesting case of $m_j = 2$, where first and second moment matching is straightforward to calculate, the required derivative $d\mathbf{V_N}(\hat{s})/d\hat{s}$ can be easily obtained, because in this case we have [169],

$$d\mathbf{M}_{\mathbf{R}}^{(\mathbf{\Omega})}(\hat{s})/d\hat{s}\Big|_{\mu_j} = df_R^*(\hat{s})/d\hat{s}|_{\mu_j}\mathbf{M}_{\mathbf{R}}^{(\mathbf{\Omega})}(\mu_j)\,\mathbf{Q_M}\mathbf{M}_{\mathbf{R}}^{(\mathbf{\Omega})}(\mu_j) \tag{A.16}$$

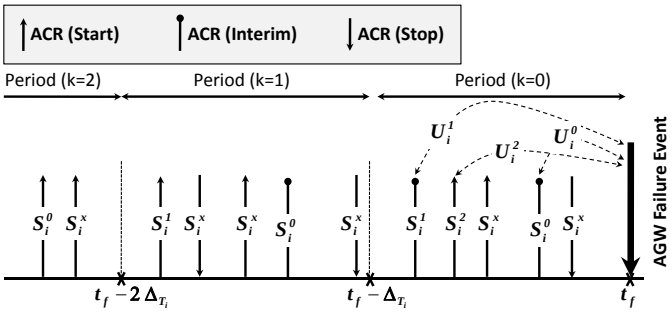


Figure A.3: The unreported usage at the event of NAS failure.

A.2 Proofs from Chapter 4

A.2.1 Proof of (4.7)

Consider a NAS failure event which occurs at an instant denoted as t_f (see Figure A.3). Let us denote the mean number of the users at the system who consume service i as N_i. When a NAS fails, the loss for service i, L_i is given by the product of the number of the users, the service cost, and the mean unreported service usage, U_i, from all users as,

$$L_i = N_i C_i U_i \tag{A.17}$$

To estimate N_i and U_i at t_f, we start by dividing the time access into Δ_{T_i} steps and move backwards from the loss event (see Figure A.3). By dividing the time axis this way, we can categorize sessions according to the number of interims they incurred (i.e., sessions with zero interims, with only one interim, etc). For instance in Figure A.3, the lifetime of session S_i^0 is less than Δ_{T_i} and hence produced no interims at the moment of failure. The unreported usage in this case is U_i^0 which equals the session lifetime. The age of the session S_i^1 at t_f lies in the interval $[\Delta_{T_i}, 2\Delta_{T_i}]$ and hence contains one interim message. The Unreported usage in this case is U_i^1. Finally, the age of session S_2 at t_f lies in the interval $[2\Delta_T, 3\Delta_T]$ and results in unreported usage of U_i^2. Any other sessions that finished before the failure event do not contribute to the loss and are marked as S_i^x in Figure A.3. In all of our exemplary cases S_i^0, S_i^1, and S_i^2, we observe that the loss event always falls randomly in the interval $[k\Delta_{T_i}, (k+1)\Delta_{T_i}]$ where $k \in \{0,1,2,\ldots\}$. Let us start by considering the sessions initiating in the first interim period such as S_2 (i.e., Period k=0) in Figure A.3. Assuming Poissonian arrivals, then the number of the corresponding sessions in the system denoted as N_i is given by the sum of the likelihood that a failure happens at instant t, a session arrival occurs (λdt), and that the session

survives until the failure event ($\bar{F}_{s_i}(t)$) as,

$$N_i^0 = \int_{t=0}^{\Delta_{T_i}} \Pr\{\text{arrival}\}\Pr\{\text{failure}\}\Pr\{\text{Session survives until t}\} = \frac{1}{\Delta_{T_i}} \int_{t=0}^{\Delta_{T_i}} \lambda_i \bar{F}_{s_i}(t)dt \tag{A.18}$$

The corresponding mean unreported usage U_i^0 *per session* is given by the weighted sum of the unreported usage due to each session divided by the number of the impacted sessions N_i^0. This is given as,

$$U_i^0 = \frac{1}{N_i^0} \frac{1}{\Delta_{T_i}} \int_{t=0}^{\Delta_{T_i}} \lambda_i t \bar{F}_{s_i}(t)dt = \frac{\int_{t=0}^{\Delta_{T_i}} t\bar{F}_{s_i}(t)dt}{\int_{t=0}^{\Delta_{T_i}} \bar{F}_{s_i}(t)dt} \tag{A.19}$$

Observing that the mean age (or residual lifetime) $E[\tilde{S}]$ for the service session until failure is given as,

$$E[\tilde{S}] = \int_{t=0}^{\infty} t\tilde{f}_S(t)dt = \int_{t=0}^{\infty} t\left(\frac{\bar{F}_S(t)}{E_s}\right)dt = \frac{\int_{t=0}^{\infty} t\bar{F}_S(t)dt}{\int_{t=0}^{\infty} \bar{F}_S(t)dt} \tag{A.20}$$

Comparing (A.19) and (A.20), it is clear that (A.19) can be viewed as the average age of the flows that have lifetime in the period of [0, Δ_{T_i}]. We now extend this result to the periods (k=1,2,3, ...). For the period (k=1), the number of arrivals is given as $N_i^1 = \frac{1}{\Delta_{T_i}} \int_{x=\Delta_{T_i}}^{2\Delta_{T_i}} \lambda_i \bar{F}_{s_i}(x)dt$. Similarly, the number of surviving arrivals in the k^{th} period is given as $N_i^k = \frac{1}{\Delta_{T_i}} \int_{x=k\Delta_{T_i}}^{(k+1)\Delta_{T_i}} \lambda_i \bar{F}_{s_i}(x)dt$. Hence, the total number of surviving arrivals from service *i* until the loss event is given as,

$$N_i = \frac{\lambda_i}{\Delta_{T_i}} \sum_{k=0}^{\infty} \int_{x=k\Delta_{T_i}}^{(k+1)\Delta_{T_i}} \bar{F}_{s_i}(x)dx = \lambda_i E_{s_i} \tag{A.21}$$

where $\sum_{k=0}^{\infty} \int_{x=k\Delta_{T_i}}^{(k+1)\Delta_{T_i}} \bar{F}_{s_i}(x)dx = \int_{x=0}^{\infty} \bar{F}_{s_i}(x)dx$. This result also matches the steady state mean number of users in M/G/∞ systems. This is because in our case we do not consider session blocking as in practice a NAS serves a large number of concurrent sessions (i.e., at least few thousands [90]) and hence the M/G/∞ is a good approximation. Using similar analysis to (A.21), the mean unreported usage, U_i, is given by the sum of the usage from all sessions starting in the periods $[k\Delta_{T_i}, (k+1)\Delta_{T_i}]$ (e.g., U_i^0, U_i^1, U_i^2 in Figure A.3).

$$U_i = \frac{1}{N_i} \sum_{k=0}^{\infty} \int_{t=0}^{\Delta_{T_i}} \frac{\lambda}{\Delta_{T_i}} t\bar{F}_{s_i}(k\Delta_{T_i}+t)dt = \frac{1}{E_{s_i}} \sum_{k=0}^{\infty} \int_{t=0}^{\Delta_{T_i}} t\bar{F}_{s_i}(k\Delta_{T_i}+t)dt \tag{A.22}$$

Let us substitute $y = k\Delta_{T_i} + t$ in (A.22). Then, we have,

$$U_i = \frac{1}{E_{s_i}} \sum_{k=0}^{\infty} \int_{y=k\Delta_{T_i}}^{(k+1)\Delta_{T_i}} (y - k\Delta_{T_i}) \bar{F}_{s_i}(y) dy \tag{A.23}$$

Observing that $\frac{\bar{F}_{s_i}(y)}{E_{s_i}}$ is simply the probability density function of the age of the service session $\tilde{S}_i$ at any random moment, (see (A.20)), then,

$$\frac{1}{E_{s_i}} \sum_{k=0}^{\infty} \int_{y=k\Delta_{T_i}}^{(k+1)\Delta_{T_i}} y \bar{F}_{s_i}(y) dy = \int_{y=0}^{\infty} y \frac{\bar{F}_{s_i}(y)}{E_{s_i}} dy = E\{\tilde{S}_i\} \tag{A.24}$$

For the other part, of (A.23) (i.e., $-\sum_{k=0}^{\infty} \int_{y=k\Delta_{T_i}}^{(k+1)\Delta_{T_i}} \frac{k\Delta_{T_i}}{E_{s_i}} \bar{F}_{s_i}(y) dy$), we observe that $\frac{\bar{F}_{s_i}(y)}{E_{s_i}}$ represents the probability density of the age of the session at t_f as $f_{\tilde{S}_i}(y)$. Using (A.1)-(A.2), a considerable simplification is made as,

$$\begin{aligned} &= -\Delta_{T_i} \sum_{k=0}^{\infty} k \int_{y=k\Delta_{T_i}}^{(k+1)\Delta_{T_i}} f_{\tilde{S}_i}(y) dy = -\Delta_{T_i} \sum_{k=0}^{\infty} k \left(\bar{F}_{\tilde{S}_i}(k\Delta_{T_i}) - \bar{F}_{\tilde{S}_i}((k+1)\Delta_{T_i}) \right) \\ &= -\Delta_{T_i} \sum_{k=1}^{\infty} \bar{F}_{\tilde{S}_i}(k\Delta_{T_i}) = -\Delta_{T_i} E\{\frac{\tilde{S}_i}{\Delta_{T_i}}\} \end{aligned} \tag{A.25}$$

Using (A.24) and (A.25), the mean unreported usage per session in (A.23) is given as,

$$U_i = E\{\tilde{S}_i\} - \Delta_{T_i} E\{\lfloor \frac{\tilde{S}_i}{\Delta_{T_i}} \rfloor\} = \varepsilon_i \Delta_{T_i} \ , \Delta_{T_i} \leq E_{s_i} \tag{A.26}$$

The upper bound on U_i can be obtained by the observation that if all sessions at the failure instant, t_f, incur at least one interim then the failure event falls uniformly in the interval $[0, \Delta_{T_i}]$ and hence the mean unreported usage per session is $\frac{\Delta_{T_i}}{2}$.

A.2.2 The Derivation of the SCLP

In this policy, we find Δ_{T_i} for each service by solving for the case when the loss constraint is bounding (i.e., $L = L_{\max}^{(j)}$) for each NAS j. To simplify the notation we drop the NAS index for the loss and the interim intervals. The interim intervals can be found by solving a linear vector equation of the steepest gradient decent direction towards the loss constraint for each NAS.

$$\boldsymbol{\Delta}_T = \boldsymbol{\Delta}_{\mathbf{T}}^{\min} - \alpha \nabla \mathbf{L} \tag{A.27}$$

The gradient function $\nabla \mathbf{L}$ for NAS j, is given by the partial derivative of the loss relative to all interim intervals served by that NAS (i.e., Δ_T) as,

$$\nabla \mathbf{L} = \left(\begin{array}{cccc} \frac{dL}{d\Delta_{T_0}} & \frac{dL}{d\Delta_{T_1}} & \cdots & \frac{dL}{d\Delta_{T_i}} \end{array} \right) \tag{A.28}$$

where $\frac{dL}{d\Delta_{T_i}} = 0.5\lambda_i C_i E_{s_i}$. Since, at the loss boundary we have $L = L_{\max}^{(j)}$, the scalar constant α is obtained substituting (A.27) into (A.26) and solve for α as,

$$\begin{aligned} L_{\max}^{(j)} &= \sum_{i \in \mathbb{N}_j} \frac{\lambda_i C_i}{2} E_{s_i} \left(\|\boldsymbol{\Delta}_{\mathbf{T}}^{\min}\|_i - \alpha \|\nabla L\|_i \right) \\ \alpha &= \frac{\sum_{i \in \mathbb{N}_j} \lambda_i C_i E_{s_i} \|\boldsymbol{\Delta}_{\mathbf{T}}^{\min}\|_i - 2L_{\max}^{(i)}}{\sum_{i \in \mathbb{N}_j} \lambda_i C_i E_{s_i} \|\nabla L\|_i} \end{aligned} \tag{A.29}$$

Appendix B Analytical Background

B.1 Transient Markov Chains

In our context, a transient Markov chain contains transient and absorbing states. For a transient state, there is a chance of never being revisited after the initial visit. This is due to the existence of absorbing states that are never abandoned once entered. We formulate the transition probabilities matrix **P** as,

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{A} \\ \mathbf{0} & \mathbf{e} \end{pmatrix},$$

where **Q** contains only transitions between transient states, **A** contains only transitions to the absorbing states Λ_i, and **e** is the identity matrix with proper dimensions.

Let the row vector **F** denote the initial probabilities for the transient chain and let $\mathbf{A_i}$ denote the i^{th} column of **A**. Then the joint probability of being absorbed into state Λ_i, which corresponds to the column $\mathbf{A_i}$, after the j^{th} transition is given as,

$$P\{N_{(i)} = j\} = \mathbf{F}\,\mathbf{Q}^{j-1}\,\mathbf{A_i}, j = 1,2,... \tag{B.1}$$

Using the fundamental matrix **M**, given as $\mathbf{M} = (\mathbf{e} - \mathbf{Q})^{-1}$, the mean number of visits to transient states, excluding the first and before absorption is [120],

$$E\{N\} = \mathbf{FMo} - 1, \tag{B.2}$$

where **o** is an all ones column vector. The -1 in (B.2) is subtracted because **FMo** includes the initial visit. The probability of being absorbed into state Λ_i is denoted as β_i and is given as [120],

$$\beta_i = \mathbf{FM}\|\mathbf{A}\|_i \text{ where } \|\mathbf{A}\|_i \text{ is the } i^{\text{th}}\text{column of } \mathbf{A} \tag{B.3}$$

B.2 The Gamma Functions and Their Properties

The incomplete gamma function falls under two categories: the lower and the upper incomplete gamma functions denoted as $\gamma(k,x)$ and $\Gamma(k,x)$ respectively. These functions

are defined by the following definite integrals,

$$\begin{aligned}\gamma(k,x) &= \int_0^x t^{k-1}e^{-t}dt \\ \Gamma(k,x) &= \int_x^\infty t^{k-1}e^{-t}dt = \Gamma(k) - \gamma(k,x) \\ \Gamma(k) &= \Gamma(k,0) = \int_0^\infty t^{k-1}e^{-t}dt \end{aligned} \tag{B.4}$$

The Laplace transforms of such functions are easily obtained from the Laplace transform of the Gamma distribution as,

$$\begin{aligned}\mathscr{L}\{\gamma(k,x)\} &= \Gamma(k)\frac{(1+\hat{s})^{-k}}{\hat{s}} \\ \mathscr{L}\{\Gamma(k,x)\} &= \Gamma(k)\frac{1-(1+\hat{s})^{-k}}{\hat{s}} \end{aligned} \tag{B.5}$$

The integration of the upper incomplete gamma function is given as,

$$\int \Gamma\left(a,\frac{x}{b}\right) = x\Gamma\left(a,\frac{x}{b}\right) - b\Gamma\left(a+1,\frac{x}{b}\right) \tag{B.6}$$

A useful relationship used for many derivations in our work, which can be obtained using integration by parts and incorporating results from (B.6), is

$$\begin{aligned}\int x\Gamma\left(a,\frac{x}{b}\right)e^{\frac{-x}{c}} &= -ce^{-\frac{x}{c}}(c+x)\Gamma\left(a,\frac{x}{b}\right) \\ &\quad + \frac{c^2\left(\frac{x}{b}\right)^a\left(\frac{x}{d}\right)^{-a}\left((b+c)\Gamma\left(a,\frac{x}{d}\right)+b\Gamma\left(a+1,\frac{x}{d}\right)\right)}{b+c} \\ d &= bc(b+c)^{-1} \end{aligned} \tag{B.7}$$

Finally, the Gamma probability density function and its Laplace transforms are,

$$\begin{aligned}\mathrm{PDF}_\gamma\left(a,\frac{x}{b}\right) &= \frac{x^{a-1}e^{\frac{-x}{b}}}{b^a\Gamma(a)} \\ \mathscr{L}\left\{\mathrm{PDF}_\gamma\left(a,\frac{x}{b}\right)\right\} &= \left(\frac{1}{1+b\hat{s}}\right)^a \end{aligned} \tag{B.8}$$

Appendix C List of Abbreviations

3GPP the Third Generation Partnership Project

3GPP2 the Third Generation Partnership Project II

AA Authentication and Authorization

AAA Authentication, Authorization, and Accounting

AGW Access Gateway

AltPPP Alternative Point-to-Point

APWC Adaptive Policy with Weight Control

ASN-GW Access Serving Node Gateway

BRPC Backward-Recursive PCE-Based Computation

BSS Business Support System

BTS Base Transceiver Station

CCDF Complementary Commutative Distribution Function

CCMP Counter Mode with Cipher Block Chaining Message Authentication Code Protocol

CDF Commutative Distribution Function

CHAP Challenge-Handshake Authentication Protocol

CSCF Call Session Control Function

CLP Constrained Loss Policy

EAP Extensible Authentication Protocol

EAP-IKE EAP-Internet Key Exchange

EAP-TLS EAP-Transport Layer Security

EAP-TTLS EAP-Tunneled Transport Layer Security

EVDO EVolution Data Optimized

GMAP Gateway Mesh Access Point

GGSN Gateway GPRS Support Node

IdP Identity Provider

IMAP Intermediate Mesh Access Point

IMS IP Multimedia Subsystem

IMSI International Mobile Subscriber Identity

i.i.d independent and identically distributed
LTE Long Term Evolution
MAP Mesh Access Point
MVNO Mobile Virtual Network Operator
NAS Network Access Server
PCE Path Computation Element
PCRF Policy and Charging Rules Function
PDF Probability Density Function
PDN-GW Packet Data Network Gateway
PDSN Packet Data Serving Node
PER Packet Error Rate
PGF Probability Generating Function
RADIUS Remote Authentication Dial In User Service
RAN Radio Access Network
RLP Radio Link Protocol
RNC Radio Network Controller
RSVP ReSerVation Protocol
SBBC Service Based Bearer Control
SCLP Simplified Constrained Loss Policy
SCTP Stream Control Transmission Protocol
S-GW Serving Gateways
SIP Session Initiation Protocol
SLA Service Level Agreement
SSO Single Sign On
TAL Total Air link Load
TCP Transport Control Protocol
TKIP Temporal Key Integrity Protocol
TO Time out
UDP User Datagram Protocol
UDR Usage Detail Record
VSPT Virtual Shortest Path Tree
VSA Vendor Specific Attribute
WAP Wireless Application Protocol
WMN Wireless Mesh Networks

Bibliography

[1] D. P. Heyman and D. Lucantoni. Modeling Multiple IP Traffic Streams with Rate Limits. IEEE/ACM Transactions on Networking, 11(6):948–958, Dec 2003.

[2] S. Bali and V. S. Frost. An Algorithm for Fitting MMPP to IP Traffic Traces. IEEE Communications Letters, 11(2):207–209, Feb 2007.

[3] J. Yasong, S. Bali, T. E. Duncan, and V. S. Frost. Predicting Properties of Congestion Events for a Queueing System With fBm Traffic. IEEE/ACM Transactions on Networking, 15(5):1098–1108, Oct 2007.

[4] Bridgewater Systems. Is Your AAA up to the WiMAX Challange ? In White Paper, Apr 2009. Available from: `http://www.bridgewatersystems.com/Assets/Downloads/Whitepapers/Is_Your_AAA_Up_To_The_WiMAX_Challenge.pdf`.

[5] J. Balbas, S. Rommer, and J. Stenfelt. Policy and Charging Control in The Evolved Packet System. IEEE Communications Magazine, 47(2):68–74, Feb 2009.

[6] P. Calhoun, M. Beadles, and A. Ratcliff. RADIUS Accounting Interim Accounting Record Extension (Draft). Jan 1998. Available from: `http://freeradius.org/rfc/draft-ietf-radius-acct-interim-01.txt`.

[7] D. Nelson and A. DeKok. Common Remote Authentication Dial In User Service (RADIUS) Implementation Issues and Suggested Fixes (RFC5080). Dec 2007.

[8] A. Dutta, V. Fajardo, R. Lopez, Y. Ohba, K. Taniuchi, and H. Schulzrinne. Media-Independent Pre-Authentication (MPA) Implementation Results (Internet Draft) . Jul 2007. Available from: `http://tools.ietf.org/html/draft-ohba-mobopts-mpa-implementation-04`.

[9] V. Y. H. Kueh and M. Wilson. Evolution of Policy Control and Charging (PCC) Architecture for 3GPP Evolved System Architecture. In Proc. of the 63rd IEEE Spring Vehicular Technology Conference (VTC'06), volume 1, pages 259–263, Melbourne, May 2006.

[10] M. Kim, M. Kim, and Y. Mun. A Hierarchical Authentication Scheme for MIPv6 Node with Local Movement Property, volume 3480. Springer Berlin/Heidelberg, May 2005.

[11] H. Moustafa, G. Bourdon, and Y. Gourhant. Authentication, Authorization and Accounting (AAA) in Hybrid Ad Hoc Hotspot's Environments. In Proc. of the 4th ACM international workshop on Wireless Mobile Applications and Services on WLAN Hotspots (WMASH '06), pages 37–46, Los Angeles, California, USA, Sep 2006.

[12] C. Ming-Chin and F. L. Jeng. LMAM: A Lightweight Mutual Authentication Mechanism for Network Mobility in Vehicular Networks. In Proc. of the IEEE Asia-Pacific Services Computing Conference (APSCC '08), pages 1611–1616, Yilan, Taiwan, 9–12 Dec 2008.

[13] P. Pereira, S. Zaghloul, A. Jukan, and S. Glaeser. Towards Innovative Vehicle to Application Server Communications: An IMS Centric Approach. In Proc. of the 4th ACM International Workshop on Mobility in the Evolving Internet Architecture (MobiArch'09), Cracow, Poland, Jun 2009.

[14] Z. Yihong, W. Dapeng, and S.M. Nettles. Authentication, Authorization, and Accounting Real-Time Secondary Market Services. In Proc. of the IEEE International Conference on Communications (ICC'05), volume 2, pages 1005–1009, Seoul Korea, 16–20 May 2005.

[15] S. Greco-Polito, M. Chamania, and A. Jukan. Extending the PCE Framework for Authentication and Authorization in Multi-domain GMPLS Networks. In IEEE International Conference on Communications (ICC'09), Dresden, Germany, Jun 2009.

[16] Bell Air Networks. Available from: `http://www.belairnetworks.com`.

[17] Y. Fang, I. Chlamtac, and Y. B. Lin. Modeling PCS Networks Under General Call Holding Time and Cell Residence Time Distributions. IEEE/ACM Transactions on Networking, 5(6):893–906, Dec 1997.

[18] P. V. Orlik and S. S. Rappaport. A Model for Teletraffic Performance and Channel Holding Time Characterization in Wireless Cellular Communication with General Session and Dwell Time Distributions. IEEE Journal on Selected Areas in Communications, 16(5):788–803, Jun 1998.

[19] S. Zaghloul, A. Jukan, and W. Alanqar. Extending QoS from Radio Access to an All-IP Core in 3G Networks: An Operator's Perspective. IEEE Communications Magazine, 45(9):124–132, Sep 2007.

[20] S. Zaghloul, W. Bziuk, and A. Jukan. A Scalable Billing Architecture for Future Wireless Mesh Backhauls. In Proc. of the IEEE International Conference on Communications (ICC '08), pages 2974–2978, Beijing, China, May 2008.

[21] Bridgewater Systems. Does Your Network Access Control Have What It Takes ? In Bridgewater Systems, 2007. Available from: http://www.bridgewatersystems.com/Assets/Downloads/Whitepapers/Does%20Your%20Network%20Access%20Control%20Have%20What%20It%20Takes%20White%20Paper.pdf.

[22] G. Camarillo and M. Garcia-Martin. The 3G IP Multimedia Subsystem (IMS). Number ISBN-10: 0470871563. John Wiley&Sons, Aug 2004.

[23] E. Andrews. Migrating to Flatter, All-IP Wireless Networks. Converge Network Digest Online Magzine, Jan 2008. Available from: http://www.convergedigest.com/bp/bp1.asp?ID=502.

[24] X. Yang and J. Bigham. An Integration Architecture to 4TH Generation Wireless Networks. In Proc. of the International Conference on Software in Telecommunications and Computer Networks (SoftCOM'06), pages 257–261, Dubrovnik, Croatia, Sep 2006.

[25] B. Raman, S. Agarwal, Y. Chen, M. Caesar, W. Cui, P. Johansson, K. Lai, T. Lavian, S. Machiraju, Z. M. Mao, G. Porter, T. Roscoe, M. Seshadri, J. S. Shih, K. Sklower, L. Subramanian, T. Suzuki, S. Zhuang, A. D. Joseph, R. H. Katz, and I. Stoica. The SAHARA Model for Service Composition across Multiple Providers, volume 2414. Springer-Verlag, 2002.

[26] Driving Mobile Content Business Revenues for Mobile Virtual Network Operators. In White Paper, Motricity, 2005. Available from: http://www.motricity.com/pdf/mvno/Motricity_MVNO_WhitePaper_Nov2005.pdf.

[27] P. Clarke. Mobile M2M Market Set for 24% CAGR Growth, Says Analyst. In EE Times Europe, July 2008. Available from: http://www.mobilehandsetdesignline.com/news/209600382.

[28] S. Cheng, K. Chen, and P. Lin. Performance Modeling on Handover Latency in Mobile IP Regional Registration. In Proc. of the 19th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'08), pages 1–5, France, 15–18 Sep 2008.

[29] Li W., Mei S., Junde S., and Lei W. An Optimal Management Domain Deployment Scheme of Hierarchical AAA in Mobile IPv6 Networks. In Proc. Canadian Conference on Electrical and Computer Engineering (CCECE'08), pages 000351–000354, Ontario, Canada, 4–7 May 2008.

[30] L. Wang, M. Song, Y. Man, Y. Zhang, J. Wang, and J. Song. A Novel Robust Fault Tolerant Scheme for Hierarchical AAA in Mobile Networks. In Proc. of the 11th International Conference on Advanced Communication Technology (ICACT'09), volume 01, pages 498–502, Gangwon-Do, Korea, 15–18 Feb 2009.

[31] A. Mishra, S. Min Ho, N. L. Petroni, T. C. Clancy, and W. A. Arbaugh. Proactive Key Distribution Using Neighbor Graphs. IEEE Wireless Communications Magazine, 11(1):26–36, Feb 2004.

[32] Y. Ohba and R. Lopez. An Extended AAA Authorization Framework with Delegation (Internet Draft). Sep 2004. Available from: `http://tools.ietf.org/html/draft-ohba-aaaarch-authorization-delegation-00`.

[33] A. Hess and G. Schaefer. Performance Evaluation of AAA/MobileIP Authentication. In Proc. of the 2nd Polish-German Teletraffic Symposium (PGTS'02), Gdansk, Poland, Sep 2002.

[34] F. McEvoy, I. Ganchev, and M. O'Droma. New Third-Party AAA Architecture and Diameter Application for 4GWW. In Proc. of the 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'05), volume 3, pages 1984–1988, Berlin, Germany, 11–14 Sep 2005.

[35] F. Eyermann, P. Racz, B. Stiller, C. Schaefer, and T. Walter. Diameter-Based Accounting Management for Wireless Services. In Proc. of the IEEE Wireless Communications and Networking Conference (WCNC'06), volume 4, pages 2305–2311, Las Vegas, Nevada, 3–6 Apr 2006.

[36] F. Yuguang. Modeling and Performance Analysis for Wireless Mobile Networks: A New Analytical Approach. IEEE/ACM Transactions on Networking, 13(5):989–1002, Oct 2005.

[37] K. Yeo and Chi-Hyuck J. Modeling and Analysis of Hierarchical Cellular Networks with General Distributions of Call and Cell Residence Times. IEEE Transactions on Vehicular Technology, 51(6):1361–1374, Nov 2002.

[38] Motorola's UMTS: Radio Network Controller Solution. In datasheet, 2007. Available from: `http://www.motorola.com/staticfiles/Business/Products/Cellular%20Networks/HSxPA/_Images/horizon_RAN_Controller_data_sheet.pdf?localeId=252`.

[39] T. Pagtzis, R. Chakravorty, J. Crowcroft, S. Hailes, and P. Kirstein. Proactive Mobile IPv6 for Context-Aware All-IP Wireless Access Networks. In Proc. of the International Conference on Wireless Networks, Communications and Mobile Computing, volume 2, pages 1017–1022, Maui, HI, 13–16 Jun 2005.

[40] C. Rigney, S. Willens, A. Rubens, and W. Simpson. Remote Authentication Dial In User Service (RADIUS) (RFC2865). Jun 2000.

[41] C. Rigney. RADIUS Accounting (RFC2866). Jun 2000.

[42] Service Requirements for the All-IP Network (AIPN) (3GPP TS 22.258). Mar 2006. Available from: `http://3gpp.org/Specification-Groups`.

[43] Network Architecture - Stage 2. Number Part 2. WiMAX Forum. Available from: http://www.wimaxforum.org/technology/documents/.

[44] Accounting Services and 3GPP2 RADIUS VSAs (3GPP2 X.S0011-005-C). Aug 2003. Available from: http://www.3gpp2.org/Public_html/specs/tsgx.cfm.

[45] P. Calhoun, J. Loughney, E. Guttman, G. Zorn, and J. Arkko. Diameter Base Protocol (RFC3588). Sep 2003.

[46] B. Aboba and J. Vollbrecht. Proxy Chaining and Policy Implementation in Roaming (RFC 2607). Jun 1999.

[47] Managing Data within Billing, Mediation, and Rating Systems. In Sleepycat Software, 2005. Available from: http://jira.atlassian.com/secure/attachment/14076/Billing++mediation+and+rating+white+paper.pdf.

[48] OSS/BSS Reference Architecture and Its Implementation Scenario for Fulfillment. In White Paper, Nokia, TietoEnator, May 2004. Available from: http://www.nokia.com/NOKIA_COM_1/About_Nokia/Press/White_Papers/pdf_files/nokia_tietoenator_0605_net.pdf.

[49] M. B. Bella., J. Eloff, and M. Olivier. A Fraud Management System Architecture for Next-Generation Networks. Forensic Science International, 185(1-3):51 – 58, 2009.

[50] A. Lior, P. Yegani, K. Chowdhury, H. Tschofenig, and A. Pashalidis. Prepaid Extensions to Remote Authentication Dial-In User Service (RADIUS) (Internet Draft). Jul 2009. Available from: http://tools.ietf.org/html/draft-lior-radius-prepaid-extensions-16.

[51] H. Hakala, L. Mattila, J-P. Koskinen, M. Stura, and Loughney J. Diameter Credit-Control Application (RFC 4006). August 2005. Available from: http://www.faqs.org/rfcs/rfc4006.html.

[52] B. Emmerson. Nokia Siemens Launches Unified Charging and Billing. In IP Communications [Online Magazine], Feb 2009. Available from: http://ipcommunications.tmcnet.com/topics/ip-communications/articles/50919-nokia-siemens-launches-unified-charging-billing.htm.

[53] Nokia Siemens Networks. Sustainable Revenue Growth. In white paper, 2009. Available from: http://w3.nokiasiemensnetworks.com/NR/rdonlyres/41383BC9-2963-42CE-AB5C-1F79F5603397/0/Sustainable_revenue_growth__the_OSS_BSS_perspective.pdf.

[54] J. Altmann and L. Rhodes. Dynamic Netvalue Analyzer - A Pricing Plan Modeling Tool for ISPs Using Actual Network Usage Data. In IEEE International Workshop on Advance Issues of E-Commerce and Web-Based Information Systems (WECWIS'02), page 153, Washington, DC, USA, Jun 2002.

[55] Nokia Siemens Networks. Unified Charging and Billing. In White Paper, 2009. Available from: http://www.nokiasiemensnetworks.com/portfolio/solutions/unified-charging-and-billing.

[56] S. Mizikovsky, Z. Wang, and H. Zhu. CDMA 1xEVDO Security. Bell Labs Tech. Journal, 11(4):291–305, 2007.

[57] Interoperability Specification (IOS) for High Rate Packet Data (HRPD) Access Network Interfaces (3GPP2 A.S0008-B). Number v1, 2006. Available from: http://3gpp.org/Specification-Groups.

[58] Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 6: Medium Access Control (MAC) Security Enhancements. In 802.11i IEEE Standard for Information technology, Telecommunications and information, exchange between systems, Local and metropolitan area networks, Specific requirements, Jul 2004.

[59] C. Perkins. IP Mobility Support for IPv4 RFC(3344). Aug 2002.

[60] C. Perkins and P. Calhoun. Authentication, Authorization, and Accounting (AAA), Registration Keys for Mobile IPv4 (RFC 3957). Mar 2005.

[61] P. Calhoun, T. Johansson, C. Perkins, T. Hiller, and P. McCann. Diameter Mobile IPv4 Application (RFC 4004). Aug 2005.

[62] S. Gundavelli, K. Leung, K. Chowdhury, and B. Patil. Proxy Mobile IPv6 (RFC 5213). Aug 2008.

[63] J. Korhonen, J. Bournelle, K. Chowdhury, A. Muhanna, and U. Meyer. Diameter Proxy Mobile IPv6: Mobile Access Gateway and Local Mobility Anchor Interaction with Diameter Server (Internet Draft). Apr 2009. Available from: http://www.h-online.com/nettools/rfc/drafts/draft-ietf-dime-pmip6-04.shtml.

[64] M. Liebsch, P. Loureiro, and J. Korhonen. Local Mobility Anchor Resolution for PMIPv6 (Internet Draft). Beijing, China, Mar 2009.

[65] Cisco Content Services Gateway Installation and Configuration Guide (R3.1). Jan 2007. Available from: http://www.cisco.com/en/US/docs/wireless/CSG/5.5/installation/configuration/guide/TD-Book-Wrapper.html.

[66] Openwave. Openwave Mobile Access Gateway. Available from: http://www.openwave.com/us/products/gateway_products/mobile_access_gateway/.

[67] 3GPP2 MMD Service Based Bearer Control (V1.0). In All-IP Core Network Multimedia Domain, Sep 2006. Available from: http://www.3gpp2.org/Public_html/specs/tsgx.cfm.

[68] G. Camarillo, W. Marshall, and J. Rosenberg. Integration of Resource Management and Session Initiation Protocol (SIP) (RFC 3312). Oct 2002.

[69] D. van Thanh, I. Jorstad, P. Engelstad, T. Jonvik, Boning F., and D. van Thuan. Authentication in a Multi-access IMS Environment. In Proc. of the IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WIMOB '08), pages 613–618, Avignon, France, 12–14 Oct 2008.

[70] D. Thanh, P. Engelstad, D. Tran, I. Jorstad, E. Bakken, D. Thuan, T. Jonvik, S. Lupetti, F. Boning, S. Millidahl, N. Bang, E. Edvardsen, and H. Kjuus. Personalised Dynamic IMS Client Using Widgets. In White Paper, 2009. Available from: http://folk.uio.no/paalee/publications/ims-thanh-gsma-2009.pdf.

[71] G. Yang, D. S. Wong, and X. Deng. Anonymous and Authenticated Key Exchange for Roaming Networks. IEEE Transactions on Wireless Communications, 6(9):3461–3472, Sep 2007.

[72] S. G. Polito, H. Schulzrinne, and A. Forte. Inter-provider AAA and Billing of VoIP Users with Token-based Method. In Proc. of the 1st International Global Information Infrastructure Symposium (GIIS'07), pages 159–166, 2–6 Jul 2007.

[73] S. Decugis. Towards a Global AAA Framework for Internet. In Proc. of the 9th Annual International Symposium on Applications and the Internet (SAINT '09), pages 227–230, Seattle, USA, Jul 2009.

[74] P. Lin, C. Shin-Ming, and L. Wanjiun. Modeling Key Caching for Mobile IP Authentication, Authorization, and Accounting (AAA) Services. IEEE Transactions on Vehicular Technology, 58(7):3596–3608, Sep 2009.

[75] D. Forsberg. Secure Distributed AAA with Domain and User Reputation. In Proc. of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'07), pages 1–6, Helsinki, Finland, 18–21 Jun 2007.

[76] H. Fathi, S. SeongHan, K. Kobara, and H. Imai. Secure AAA and Mobility for Nested Mobile Networks. In Proc. of the 7th International Conference on ITS Telecommunications (ITST '07), pages 1–6, Sophia Antipolis, France, Jun 2007.

[77] H. Haverinen, N. Asokan, and T. Maattanen. Authentication and Key Generation for Mobile IP Using GSM Authentication and Roaming. In Proc. IEEE International Conference on Communications (ICC'01), volume 8, pages 2453–2457, Helsinki, Finland, 11–14 Jun 2001.

[78] M. S. Siddiqui and S. H. Choong. Security Issues in Wireless Mesh Networks. In Proc. of the International Conference on Multimedia and Ubiquitous Engineering (MUE '07), pages 717–722, Seoul, Korea, 26–28 Apr 2007.

[79] H. Redwan and Kim Ki-Hyung. Survey of Security Requirements, Attacks and Network Integration in Wireless Mesh Networks. In Proc. of the Japan-China Joint Workshop on Frontier of Computer Science and Technology (FCST '08), pages 3–9, Nagasaki, Japan, 27–28 Dec 2008.

[80] H. Yokota, A. Idoue, T. Hasegawa, and T. Kato. Link Layer Assisted Mobile IP Fast Handoff Method over Wireless LAN Networks. In Proc. of the 8th ACM annual international conference on Mobile Computing and networking (MobiCom '02), pages 131–139, New York, NY, USA, 2002.

[81] J. Seol and J. Chung. IEEE 802.21 MIH based Handover for Next Generation Mobile Communication Systems. In Proc. of the 4th International Conference on Innovations in Information Technology Innovations, pages 431–435, 18–20 Nov 2007.

[82] V. K. Gondi, Quoc-Thinh N.-V., and N. Agoulmine. A New Mobility Solution Based On PMIP Using AAA Mobility Extensions in Heterogeneous Networks. In Proc. of the IEEE Network Operations and Management Symposium Workshops (NOMS'08), pages 39–43, 7–11 April 2008.

[83] R. Koodli. Fast Handovers for Mobile IPv6 (RFC 4068). Jul 2005.

[84] Y. Wei-Zu, L. Fang-Sun, and C. Ming-Feng. Performance Modeling of Integrated Mobile Prepaid Services. IEEE Transactions on Vehicular Technology, 56(2):899–906, Mar 2007.

[85] S. I. Sou, Y. B. Lin, Q. Wu, and J. Y. Jeng. Modeling Prepaid Application Server of VoIP and Messaging Services for UMTS. IEEE Transactions on Vehicular Technology, 56(3):1434–1441, May 2007.

[86] S. Sou, H. Hung, Y. Lin, N. Peng, and J. Jeng. Modeling Credit Reservation Procedure for UMTS Online Charging System. IEEE Transactions on Wireless Communications, 6(11):4129–4135, Nov 2007.

[87] S. Sok-ian, L. Yi-bing, and J. Jeu-yih. Reducing Credit Re-authorization Cost in UMTS Online Charging System. IEEE Transactions on Wireless Communications, 7(9):3629–3635, Sep 2008.

[88] J. Zhang, J. Li, S. Weinstein, and N. Tu. Virtual Operator Based AAA in Wireless LAN Hot Spots with Ad-Hoc Networking Support. In ACM SIGMOBILE Mobile Computing and Communications Review, volume 6, pages 10–21, New York, NY, USA, 2002.

[89] E. Coronado and S. Cherkaoui. An AAA Study for Service Provisioning in Vehicular Networks. In Proc. of the 32nd IEEE Conference on Local Computer Networks (LCN'07), pages 669–676, Clontarf Castle, Dublin, Ireland, Oct 2007.

[90] Packet Data Serving Node(PDSN)/Foreign Agent (FA) and Home Agent (HA). Available from: `http://www.starentnetworks.com/File/Starent_Networks_PDSN_0509.pdf`.

[91] S. Zaghloul and A. Jukan. On the Performance of the AAA Systems in 3G Cellular Networks. In Proc. of the IEEE International Conference on Communications (ICC '07), pages 2103–2108, Glasgow, Scotland, 24–28 Jun 2007.

[92] S. Zaghloul and A. Jukan. Relating the AAA and the Radio Access Rates in 3G Cellular Networks. IEEE Communications Letters, 11(4):363–365, Apr 2007.

[93] Radius Client-MikroTik RouterOS V2.9. Available from: `http://www.mikrotik.com/docs/ros/2.9/guide/aaaradius`.

[94] R. Nelson. Probability, Stochastic Processes, and Queuing Theory - The Mathematics of Computer Performance Modeling. Number ISBN: 0387944524. Springer, 1995.

[95] F. Barcelo and J. Jordan. Channel Holding Time Distribution in Public Telephony Systems (PAMR and PCS). IEEE Transactions on Vehicular Technology, 49(5):1615–1625, Sep 2000.

[96] E. Casilari, H. Montes, and F. Sandoval. Modelling of Voice Traffic Over IP Networks. In Proc. of the 3rd International Symposium on Communications Systems Networks and Digital Signal Processing (CSNDSP'02), pages 411–414, Staffordshire, UK, Jul 2002.

[97] C. Jedrzycki and V. C. M. Leung. Probability Distribution of Channel Holding Time in Cellular Telephony Systems. In Proc. of the 46th IEEE Vehicular Technology Conference (VTC'96), volume 1, pages 247–251, Atlanta, GA, USA, 28 April–1 May 1996.

[98] L. Peng, T. Cailin, M. Jie, C. Yongyu, and Y. Dacheng. Experimental Study on Traffic Model of Wireless Internet Services in CDMA Network. In Proc. of the 61st IEEE Vehicular Technology Conference (VTC'05), volume 4, pages 2137–2141, Stockholm, Sweden, 30 May–1 Jun 2005.

[99] E. A. Yavuz and V. C. M. Leung. Modeling Channel Occupancy Times for Voice Traffic in Cellular Networks. In Proc. of the IEEE International Conference on Communications (ICC '07), pages 332–337, Glasgow, Scotland, 24–28 Jun 2007.

[100] P. Calhoun, G. Zorn, D. Spence, and D. Mitton. Diameter Network Access Server Application (RFC4005). Aug 2005.

[101] All-IP Core Network Multimedia Domain (3GPP2 X.S0013-000-B). Dec 2007. Available from: `http://www.3gpp2.org/Public_html/specs/tsgx.cfm`.

[102] S. Zaghloul and A. Jukan. Signaling Rate and Performance for Authentication, Authorization, and Accounting (AAA) Systems in All-IP Cellular Networks. IEEE Transactions on Wireless Communications, 8(6):2960–2971, Jun 2009.

[103] W. Liang and W. Wang. On Performance Analysis of Challenge/Response Based Authentication in Wireless Networks. Elsevier Computer Networks Journal, 48:267–288, Jun 2005.

[104] A. Abdi and M. Kaveh. K Distribution: An Appropriate Substitute for Rayleigh-Lognormal Distribution in Fading-Shadowing Wireless Channels. Electronics Letters, 34(9):851–852, Apr 1998.

[105] A. Thummler, P. Buchholz, and M. Telek. A Novel Approach for Phase-Type Fitting with the EM Algorithm. IEEE Transactions on Dependable and Secure Computing, 3(3):245–258, Jul–Sep 2006.

[106] R. M. Rodriguez-Dagnino and H. Takagi. Counting Handovers in a Cellular Mobile Communication Network: Equilibrium Renewal Process Approach. Performance Evaluation, 52(2):153–174, Apr 2003.

[107] W. Bziuk, S. Zaghloul, and A. Jukan. A New Framework for Characterizing the Number of Handoffs in Cellular Networks. In The Fifth Polish-German Teletraffic Symposium (PGTS'08), Berlin, Germany, Sep 2008.

[108] W. Bziuk, S. Zaghloul, and A. Jukan. A New Framework for Characterizing the Number of Handoffs in Cellular Networks. European Transactions on Telecommunications, 20(7):689–700, Nov 2009.

[109] M. S. Sricharan and V. Vaidehi. A Pragmatic Analysis of User Mobility Patterns in Macrocellular Wireless Networks. In Proc. of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'07), pages 1–9, 18–21 Jun 2007.

[110] S. Mohanty and I. F. Akyildiz. Performance Analysis of Handoff Techniques Based on Mobile IP, TCP-Migrate, and SIP. IEEE Transactions on Mobile Computing, 6(7):731–747, Jul 2007.

[111] H. Yokota and G. Dommety. Mobile IPv6 Fast Handovers for 3G CDMA Networks (RFC 5271). Jun 2008.

[112] PPP-Alternate Protocol (AltPPP) for CDMA2000 - Wireless IP Network Standard (3GPP2 X.S0040-0). Jan 2007. Available from: `http://www.3gpp2.org/Public_html/specs/tsgx.cfm`.

[113] M. Claypool, R. Kinicki, W Lee, M. Li, and G. Ratner. Characterization By Measurement of a CDMA 1x EVDO Network. In Proc. of the 2nd ACM annual International Workshop on Wireless Internet (WICON '06), page 2, New York, NY, USA, 2006.

[114] S. Nanda. Teletraffic Models for Urban and Suburban Microcells: Cell Sizes and Handoff Rates. IEEE Transactions on Vehicular Technology, 42(4):673–682, Nov 1993.

[115] M. O'Droma and I. Ganchev. Toward A Ubiquitous Consumer Wireless World. IEEE Wireless Communications Magazine, 14(1):52–63, Feb 2007.

[116] I. F. Akyildiz, J. S. M. Ho, and L. Yi-Bing. Movement-Based Location Update and Selective Paging for PCS Networks. IEEE/ACM Transactions on Networking, 4(4):629–638, Aug 1996.

[117] I. F. Akyildiz and W. Wang. A Dynamic Location Management Scheme for Next-Generation Multitier PCS Systems. IEEE Transactions on Wireless Communications, 1(1):178–189, Jan 2002.

[118] A. Bar-Noy, I. Kessler, and M. Sidi. Mobile Users: To Update or not to Update? Wireless Networks, 1(2):175 – 185, Jun 1995. Available from: `http://www.springerlink.com/content/g5763360828v3113`.

[119] A. Bar-Noy and I. Kessler. Tracking Mobile Users in Wireless Communications Networks. In Proc. Twelfth Annual Joint Conference of the IEEE Computer and Communications Societies. Networking: Foundation for the Future (INFOCOM '93), pages 1232–1239, San Francisco, CA, USA, 28 Mar–1 Apr 1993.

[120] U. Bhat and G. Miller. Elements of Applied Stochastic Processes. Number ISBN: 978-0-471-41442-1. Wiley, 3rd edition, 2002.

[121] J. Kemeny and L. Snell. Finite Markov Chains. Number ISBN: 9780387901923. Springer, 3rd edition, 1983.

[122] IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum. IEEE, 2006.

[123] J. Loughney, M. Nakhjiri, C. Perkins, and R. Koodli. Context Transfer Protocol (CXTP) (RFC 4067). Jul 2005.

[124] S. Zaghloul, W. Bziuk, and A. Jukan. Signaling and Handoff Rates at the Policy Control Function (PCF) in IP Multimedia Subsystem (IMS). IEEE Communications Letters, 12(7):526–528, Jul 2008.

[125] W. Bziuk, S. Zaghloul, and A. Jukan. The Spatial Effect of Mobility on the Mean Number of Handoffs: A New Theoritical Result. In Proc. of the IEEE International Communications Conference (ICC'09), Dresden, Germany, Jun 2009.

[126] R. Perera, A. Eisenblatter, E. Fledderus, C. Gorg, M. Scheutzow, and S. Verwijmeren. Pixel Oriented Mobility Modelling for UMTS Network Simulations. In Proc. of the IST Mobile & Wireless Telecommunications Summit (IST-2000-28088 Momentum), Thessaloniki, Greece, Jun 2002.

[127] S. Zaghloul, W. Bziuk, and A. Jukan. A Novel Analytical Framework for Mobility Modeling in All-IP Wireless Systems. In Proc. of the 21st International Teletraffic Congress (ITC 21), Paris, France, Sep 2009.

[128] W. Bziuk, S. Zaghloul, and A. Jukan. Revisiting Handoff Statistics under General Assumptions for Sessions, Networks and Mobility. In to be submitted to the 22nd International Teletraffic Congress (ITC 22), 2010.

[129] Modules from FreeRADIUS. Available from: `http://wiki.freeradius.org/Modules`.

[130] Y. Ohba. Diameter NASREQ Application API. 2004. Available from: `http://www.opendiameter.org/`.

[131] A. Dutta, K. Manousakis, S. Das, and F. J. Lin. Mobility Testbed for 3GPP2-Based Multimedia Domain Networks. IEEE Communications Magazine, 45(7):118–126, Jul 2007.

[132] K. Taniuchi, Y. Ohba, V. Fajardo, S. Das, M. Tauil, Yuu-Heng C., A. Dutta, D. Baker, M. Yajnik, and D. Famolari. IEEE 802.21: Media Independent Handover: Features, Applicability, and Realization. IEEE Communications Magazine, 47(1):112–120, Jan 2009.

[133] L. Dimopoulou, G. Leoleis, and I.O. Venieris. Fast Handover Support in A WLAN Environment: Challenges and Perspectives. Network, IEEE, 19(3):14–20, May-Jun 2005.

[134] A. Udugama, M. Iqbal, U. Toseef, C. Goerg, C. Fan, and M. Schlaeger. Evaluation of a Network Based Mobility Management Protocol: PMIPv6. In Proc. of the 69th IEEE Vehicular Technology Conference (VTC'09), pages 1–5, Barcelona, 26–29 Apr 2009.

[135] I. Ali, A. Casati, K. Chowdhury, K. Nishida, E. Parsons, S. Schmid, and R. Vaidya. Network-Based Mobility Management in The Evolved 3GPP Core Network. IEEE Communications Magazine, 47(2):58–66, Feb 2009.

[136] S. Zaghloul, J. I. Aznar, and A. Jukan. Application Layer Signaling for Proactive Handoff Management in All-IP Wireless Networks. In Proc. of the IEEE International Conference on Communications (ICC '09), pages 1–6, Dresden, Germany, 14–18 Jun 2009.

[137] H. Ekstrom. QoS Control In The 3GPP Evolved Packet System. Communications Magazine, IEEE, 47(2):76–83, Feb 2009.

[138] Service Based Bearer Control - Ty Interface Stage-3 (3GPP2 X.S0013-014-0). In All-IP Core Network Multimedia Domain, volume Ver 1.0, Dec 2007. Available from: `http://www.3gpp2.org/Public_html/specs/tsgx.cfm`.

[139] M. Chiba, G. Dommety, M. Eklund, D. Mitton, and B. Aboba. Dynamic Authorization Extensions to Remote Authentication Dial In User Service (RADIUS) (RFC 3576). Jul 2003.

[140] J. I Aznar. A Proactive Signaling Mechanism for Service Oriented QoS Control in Future All-IP Wireless Networks. Master's thesis, Technische Universitaet Carolo-Wilhelmina zu Braunschweig, 2008.

[141] S. Qingyang and A. Jamalipour. Network Selection in An Integrated Wireless LAN and UMTS Environment Using Mathematical Modeling and Computing Techniques. IEEE Wireless Communications Magazine, 12(3):42–48, Jun 2005.

[142] V. Nair. Evolution of QoS and Charging Framework in WiMAX. 2009. Available from: `http://www.wimax.com/commentary/spotlight/evolution-of-qos-and-charging-framework-in-wimax`.

[143] J. Koomey. Estimating Total Power Consumption By Servers in The U.S. and the World. In Final Report. Lawrence Berkeley National Laboratory (LBNL), 2007. Available from: `http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf`.

[144] J. Na, Y. Chung, M. Yun, and Y. Kim. An Efficient Diameter-Based Accounting Scheme for Wireless Metropolitan Area Network (WMAN). In Prococeedings of the 60th IEEE Vehicular Technology Conference (VTC'04), volume 7, pages 5072–5076, Milan, 26–29 Sep 2004.

[145] S. Zaghloul and A. Jukan. Optimal Accounting Policies for AAA Systems in Mobile Telecommunications Networks. to appear in the IEEE Transactions on Mobile Computing, Jun 2009.

[146] US Patent 6999449 - System and Method of Monitoring and Reporting Accounting Data Based on Volume. Feb 2006.

[147] AAA Service Controller. In datasheet. Bridgewater Systems, 2007. Available from: www.bridgewatersystems.com/products/aaa_service_controller.html.

[148] User Guide for Cisco Access Registrar 4.2. In Cisco, 2008. Available from: http://www.cisco.com/en/US/docs/net_mgmt/access_registrar/4.2/user/guide/CAR4.2_usersguide.pdf.

[149] O. Tipmongkolsilp, S. Zaghloul, and A. Jukan. The Evolution of Cellular Backhaul Technologies:Current Issues and Future Trends. to appear in the IEEE Communications Surveys & Tutorials Journal, 1st Quarter 2011.

[150] K. Giesken. Application of Wireless Technology in the Mobile Backhaul Network. Bechtel Telecommunications Technical Journal, 12, 2002.

[151] Timing and synchronization in Next-Generation Wireless Networks. Symmetricom, 2008. Available from: http://ngn.symmetricom.com/pdf/application_notes/AN_MPT.pdf.

[152] J. He, K. Yang, K. Guild, and H. Chen. Application of IEEE 802.16 Mesh Networks as the Backhaul of Multihop Cellular Networks. IEEE Communications Magazine, 45(9):82–90, Sep 2007.

[153] D. Chen. On the Analysis of Using 802.16e WiMAX for Point-to-Point Wireless Backhaul. In Proc. of the IEEE Radio and Wireless Symposium, pages 507–510, Long Beach, CA, 9–11 Jan 2007.

[154] E. Carlson, C. Prehofer, C. Bettstetter, H. Karl, and A. Wolisz. A Distributed End-to-End Reservation Protocol for IEEE 802.11-Based Wireless Mesh Networks. IEEE Journal on Selected Areas in Communications, 24(11):2018–2027, Nov 2006.

[155] D. Mitton and M. Beadles. Network Access Server Requirements Next Generation (NASREQNG) (RFC 2881). Jul 2000.

[156] Q. Bi, P. Chen, Y. Yang, and Q. Zhang. An Analysis of VoIP Service Using 1xEV-DO Revision A System. IEEE Journal on Selected Areas in Communications, 24(1):36–45, Jan 2006.

[157] R. Douville, J. L. Le Roux, J. L. Rougier, and S. Secci. A Service Plane over The PCE Architecture for Automatic Multidomain Connection-Oriented Services. IEEE Communications Magazine, 46(6):94–102, Jun 2008.

[158] J-P. Vasseur, R. Zhang, N. Bitar, and J-L. Le Roux. A Backward Recursive PCE-based Computation (BRPC) Procedure To Compute Shortest Constrained Inter-domain Traffic Engineering Label Switched Paths (Internet Draft). Apr 2008. Available from: http://tools.ietf.org/id/draft-ietf-pce-brpc-09.txt.

[159] A. Farrel and J-P. Vasseur. A Path Computation Element (PCE)-Based Architecture (RFC4655). Aug 2006.

[160] J-P. Vasseur and J-L. Le Roux. Path Computation Element (PCE) Communication Protocol (PCEP) (Internet Draft). Aug. 2008. Available from: `http://www.ietf.org/internet-drafts/draft-ietf-pce-pcep-13.txt`.

[161] R. Bradford, J.-P. Vasseur, and A. Farrel. Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism (RFC 5520). Apr 2009.

[162] BRITE: Boston University Representative Internet Topology Generator. Available from: `http://www.cs.bu.edu/brite/`.

[163] D. Sun, P. McCann, H. Tschofenig, T. Tsou, A. Doria, and G. Zorn. Diameter Quality of Service Application (Internet Draft). May 2009. Available from: `http://tools.ietf.org/id/draft-ietf-dime-diameter-qos-08.txt`.

[164] G. Leibovitz. Roadmap for Cryptographic Authentication of Routing Protocol Packets on the Wire (Internet Draft). Sep 2009. Available from: `http://tools.ietf.org/id/draft-lebovitz-kmart-roadmap-02`.

[165] H. Hui-Nien, L. Pei-Chun, and L. Yi-Bing. Random Number Generation for Excess Life of Mobile User Residence Time. IEEE Transactions on Vehicular Technology, 55(3):1045–1050, May 2006.

[166] P. Funk and S. Blake-Wilson. Extensible Authentication Protocol Tunneled Transport Layer Security Authenticated Protocol Version 0 (EAP-TTLSv0) (RFC 5281). Aug 2008.

[167] J. Yoon, M. Liu, and B. Noble. Random Waypoint Considered Harmful. In Proc.of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003), volume 2, pages 1312–1321, San Francisco, Mar 2003.

[168] JDiameter Project. Available from: `https://jdiameter.dev.java.net/`.

[169] W. Fischer and K. Meier-Hellstern. The Markov-Modulated Poisson Process (MMPP) Cookbook. Performance Evaluation, 18(2):149–171, 1992.

Mobile Telecommunications Networks

Authentication, Authorization, and Accounting (AAA) systems have been and will continue to be pivotal elements for the success of mobile telecom networks. In their basic operation, AAA systems grant users the required access and facilitate the collection of accounting records which reflect the subscribers' usage of network resources. The design of AAA systems is therefore instrumental to the operators' revenue growth as it largely depends on ensuring transparent verification of users' identities, quickly authorizing the requested QoS levels by the services, and implementing smart charging and accounting strategies for the services.

In light of these developments, in this thesis, we lay the foundations for the first formal framework for AAA system planning by extending fundamental results from cellular performance studies. We also propose novel optimization mechanisms to improve accounting reliability and to mitigate authentication delay during handoffs in multi-service mobile networks. We also introduce AAA protocols to two promising areas including cellular backhaul applications over wireless mesh networks and inter-operator layer 2 optical systems.

This thesis is one of few works that specialize in the performance of AAA systems at the fundamental level. It is part of on-going work on designing next generation mobile architectures and control planes. As AAA systems continue to play a pivotal role to the success of mobile systems, further research is still necessary to bring the design of AAA systems to maturity and with tighter integration with business processes and models.

Said Zaghloul is a research staff member at the Technical University of Braunschweig and has worked with major international corporations including Sprint, USA and Siemens, Germany. During his studies, he received international awards including the Fulbright scholarship and the IET first award for BSc senior projects. Said's current research interests include: AAA systems, next generation all-IP wireless architectures, signaling plane performance, mobility, and wireless communications.

Braunschweig, Germany, 2010